Evaluation of Cross-Language Information Retrieval Using the Domain-Specific GIRT Data as Parallel German-English Corpus

Michael Kluck

Informationszentrum Sozialwissenschaften (IZ)
Lennéstr. 30, 53113 Bonn, Germany
kluck@bonn.iz-soz.de

Abstract

The development of the evaluation of domain-specific cross-language information retrieval (CLIR) is shown in the context of the Cross-Language Evaluation Forum (CLEF) campaigns from 2000 to 2003. The pre-conditions and the usable data and additionally available instruments are described. The main goals of this task of CLEF are to allow the evaluation of Cross-Language Information Retrieval (CLIR) systems in the context of structured data and in a domain-specific area (not in the more general context of floating, journalistic texts), and with the additional possibility to make use of thesauri which had been used for intellectual indexing of the documents and are provided with the data. The parallel German-English GIRT4 corpus is described and some of the results of the CLEF 2004 campaign are discussed.

Domain-Specific CLIR in the Context of CLEF

The development of the evaluation of domain-specific Cross-Language Information Retrieval (CLIR) is embedded in the context of the Cross-Language Evaluation Forum (CLEF)[1] campaigns from 2000 to 2003. The main goals of this task of CLEF are to allow the evaluation of CLIR systems in the context of structured data and in a domain-specific area (not in the more general context of floating, journalistic texts), and with the additional possibility to make use of thesauri which had been used for intellectual indexing of the documents and are provided together with the data. The general purpose of the work on GIRT within CLEF is discussed in Kluck/Gey (2001).

The data provided for this task have been GIRT (= German Indexing and Retrieval Testdatabase)[2] and Amaryllis[3]. The GIRT corpus has been used in several versions for a number of retrieval tests in Germany, in TREC[4] and CLEF. The first pre-test with GIRT data has been carried out in 1997 (see Kluck, 1998)[5]. Amaryllis has also been part of other tests in France and French speaking countries.

GIRT and Amaryllis Tasks in CLEF 2000-2002

In 2000, 2001 and 2002 the special task of CLEF on "Domain-Specific Mono- and Cross-Language Information Retrieval" used the GIRT3 corpus consisting of a data collection from a vertical domain (social sciences); this collection contained more than 76,000 documents in a structured database. This special task offered 25 queries (topics) each year[6], created in German, but also translated into English and Russian. Participating groups could run these topics:

1. either as monolingual task (German topics) against the 76,000 German documents of this database (GIRT3);
2. or as multilingual task using the translated topics.

In addition a German-English thesaurus and a German-Russian wordlist as well as English translations of the document titles have been available.

In 2002, there was an additional distinct tasks with the Amaryllis corpus to test system performance in searching a multi-disciplinary scientific database of approximately 150,000 French bibliographic documents. As additional tool a controlled vocabulary in English and French was provided that could be used in the retrieval task. Topics have been provided in English and French.

GIRT Task in CLEF 2003

In the CLEF campaign 2003 the GIRT track used a new much larger collection: GIRT4. This collection of German social science data contains 151,319 documents and is available as two pseudo-parallel corpora which contain the same documents:

* in German (GIRT4-DE) and
* in English (GIRT4-EN)

Again the topics have been provided in German, English and Russian language.

Thus, in the CLEF 2003 campaign it was possible to offer two ***monolingual tasks***:

1. using German topics against German data,
2. using English topics against English data,

and two bilingual tasks:

1. using English or Russian topics against German data,
2. using German or Russian topics against English data.

The same controlled vocabularies in German-English and German-Russian as in the previous CLEF campaigns were available.

[1] See www.clef-campaign.org and Peters et al., (2004, 2003, 2002), Peters (2001), Kluck (2002)

[2] For an extended overview on the GIRT versions and the tests carried out with these see Kluck (2003, 2004).

3 Now available via ELDA.

[4] See Vorhees/Harman (2000).

[5] The German-English Thesaurus for the Social Sciences and the German-Russian wordlist are provided by the IZ (German Social Science Information Centre) as machine readable files. These files have been extracted from the following printed thesauri: Schott (1999) and Basarnova et al. (1997)

[6] On the topic creation process see Kluck/Womser-Hacker (2002)

Structure of the Parallel GIRT4 Corpora

As the GIRT4-DE and GIRT4-EN data are in parallel, most information elements are existing in both corpora. But due to the fact that the English corpus is a translation of the original German one, some information elements (data entry fields) are not available to the same extent.

Information element (field)	Number of entries, all documents	Number of entries in this field per document in GIRT4-DE	Number of entries in this field per document in GIRT4-EN
Document no. (DE = EN)	151,319	1	1
Author (DE = EN)	237,301	1.75	1.75
Title (DE = EN)	151,319	1	1
Key words (DE = EN)	1,535,709	10.15	10.15
Classification text (DE = EN)	305,504	2.02	2.02
Methodologi-cal key words (DE)	354,968	2.35	-
Methodologi-cal key words (EN)	292,387	-	1.93
Abstract (DE)	145,941	0.96	-
Abstract (EN)[7]	22,058	-	0.15
Free terms (only DE)	38,505	0.25	-
Methodologi-cal text (only DE)	10,258	0.07	-

Figure 1: Distribution of fields in the German and the English Part of GIRT4

As it was the condition for the extraction of the documents from the source databases SOLIS and FORIS[8] each document in both corpora has a title field. The translation of the German title has been done by human translators. Each document carries a document number which is the information element that identifies them. We have randomly changed the document number in the English part to make it not too easy to identify corresponding documents[9].

The documents have in most cases more than one author (on average 1.75). Each document has intellectually been indexed by about 10 key words. The indexing was done in German, the English equivalents have been taken form the German-English thesaurus (Schott 1998). The same has been done with classification texts which occur on average twice per document. About one or two methodological key words have been assigned per document with slight differences between the German and the English part which are caused by the reduced term list in the English part. Free terms and methodological texts are only offered in the German part, and are assigned to 25% or 7% of the documents. For 96% of all German documents abstracts are provided, whereas only 15% are available for the English part. This is the main reason for the reduced size of the English corpus and why we call this a pseudo-parallel corpus.

Relevance Assessment of the Parallel Corpora

The documents of the two parallel corpora have differently been numbered, thus, the participating groups could not automatically detect which of the documents have been the same (but only translated from German into English).

For measuring the relative performance of the retrieval systems we applied the pooling method developed by the TREC initiative (Vorhees/Harman, 2000) The systems participating in this track delivered the top 60 results for each topic which they suspected to be the most relevant answers to the given query (topic). The results were grouped by topic. These topic related lists of documents were then judged by human assessors. The relevance assessment was done in a binary way: a document was either counted as relevant with respect to the topic in question or not.

Then, we have reconstructed the concordance of the numbering in both corpora (and concatenated the identical documents). Thus, we could make an in depth comparison of the results (compare Figure 2).

During the CLEF 2004 campaign a total of 17,031 documents from GIRT4-DE and GIRT4-EN was delivered as relevant hits by the participating groups. These formed the pool of documents to be intellectually assessed. As 25 topics have been used in this campaign, on average 681 documents had to be assessed per topic. In the end, out of these suspected relevant hits 3,449 or 20.25% have been judged as really relevant.

Differences in the Number of Result Hits

A general observation is that the hits of relevant documents in the result sets from the both corpora are not fully identical (which would have been the optimal outcome). There have been 11,137 hits delivered from GIRT4-DE and 5,894 from GIRT4-EN. Within the German result hits 8,993 or 80.75% did not have the corresponding English document. And within the English results 3,756 or 63.7% did not have the corresponding German document. That means only 2.138 document have been included in both sets and therefore had to be judged twice. All in all this overlap in the results of the German and English part is quite low. But there is no evidence of any significant correlation to specific topical queries. And it must be considered that the majority of runs delivered by the participants has been aimed on the German collection. This fact obviously caused a predominance of German results which were not accompanied by respective results in the English part.

Differences in Assessments

In few cases there was a different judgment done by the assessor for the same document with respect to the same topic in the German and English collection. This occurred in 171 cases or 1% of all judgments. There was no significant correlation to one specific topic, only two topics had not been touched by this problem. On average 7

[7] The translation of German abstract has partly been done by human translators (HT) and partly by machine translation (MT). The MT was done by SYSTRAN.

[8] The databases SOLIS (Social Science Literature Information System) and FORIS (Social Science Research Information System) are produced and provided by the IZ, Bonn.

[9] A list of the corresponding document numbers of the German and the English part is available at the IZ, but was not forwarded to the participants of the CLEF campaign.

cases of unequal judgments occurred per topic with the highest value of 17 cases and the lowest of 1 case.

The re-assessment of these 171 documents resulted in 57% of the cases in a change from irrelevant to relevant, and in 43% in a change from relevant to irrelevant which is nearly the same amount. Overall the re-assessment indicated 98 documents as relevant. But the changes have a little bit more been related to the English part of the corpus (60%). This observation and the fact that less results have been delivered from the English part (only 34,61% of all results) emphasizes the assumption that the reduced extent of text in the English part made judgments more difficult or vague (because of the lack of extended information which is mainly carried by the abstracts).

	GIRT4-DE		**GIRT-EN**		**sum**	
	n	**%**	**n**	**%**	**n**	**%**
Unique documents in total	151.319	100	151.319	100	-	-
Unique documents assessed	-	-	-	-	13.412	8,86
All assessments done (with overlap)	11.132	7,29	5.893	3,86	17.025	-
Unique documents assessed as relevant					3.442	
Documents differently assessed (by unique assessments done)	-	-	-	-	171	1,00 *(of n= 17.025)*
Unique assessed documents not in the parallel corpus	8.993		3.756		12.749	
Re-assessed documents relevant					97	56,73 *(of n = 171)*

Figure 2: Delivered Hits in CLEF 2004 for GIRT4-DE and GIRT4-EN

GIRT Task Participants in the CLEF 2004 Campaign

In 2004 four groups participated in the GIRT Task: University of California at Berkeley (USA), Distance University Hagen (Germany), ENEA/University La Sapienza Rome[10] (Italy), University Amsterdam (Netherlands).

The University Amsterdam, which belongs to the leading groups in nearly all monolingual and multilingual tasks of CLEF, used a vector space model with 100-dimensional space (Kamps et al., 2003). Further they used a stemmer for the re-ranking (but without decomposition of complex words), and alternatively a 4-gramm-model. The n-gramm-method (here with n=4) searches character strings which are n characters long and identifies identical character strings[11]. By that, this method does not need any knowledge of the respective languages. Additional improvements could be made by using the indexing of the provided documents.

The University of California at Berkeley (Petras/ Perelman/Gey 2003), which participate in all TREC and CLEF campaigns since the beginning and always has been belonging to the best performing groups in all tasks, has made use of all GIRT4 sub-tasks. They clearly showed, that the use of thesauri lead to a remarkable improvement of results, although the publicly available machine translation systems (MT) have reached a better quality meanwhile. The best results have been achieved by the combination of two MT systems with the usage of the thesaurus: “Documents that have controlled vocabulary terms added to the usual title and abstract information prove advantageous in retrieval because the thesaurus terms add valuable search terms to the index. An index containing titles, abstracts, and thesaurus terms will always outperform an index only containing title and abstract.”[12]

ENEA/University Rome La Sapienza (Alderuccio/ Bordoni/Loreto 2003) have chosen a totally different approach than the usual CLIR systems, namely the data compression, which should them enable, to detect the syntactical and semantic distance of character strings, without having any knowledge of the respective languages and their peculiarities.

The Distance University Hagen (Leveling 2003) introduced another approach into the CLIR evaluation in CLEF, which is based on a natural language interface. To analyze the texts of the topics and the documents, multiple lexical and morphological information and resource were used, especially those supporting the disambiguation of meanings of single character strings and the decomposition of compounds. For producing a searchable database of the GIRT data they used the Zebra software, which provides a Z39.50 interface and relevance operator and allows ranking of results. The Social Science Thesaurus has also been used and provided as a lexical resource in MultiNet manner.

Conclusion

For now the GIRT4 data have been a valuable source for CLIR evaluation, although not yet all possible facets have been exploited. But these data also offer chances as a source for linguistic research as they give a lot of real parallel texts (titles and indexing terms, and as far as abstracts exist) in two languages. They may also be useful to determine co-occurrences of intellectually assigned indexing terms and terms in the free text. We hope to enlarge the domain-specific task of CLEF by adding other English, Russian, and French corpora. Then a complete multilingual sub-task would be possible and comparable corpora would by available to the scientific community.

[10] ENEA = Ente per le Nuove tecnologie, l’Energia e l’Ambiente, S. Maria di Galeria (Roma); Università degli Studi di Roma La Sapienza

[11] For instance the phrase „information retrieval“ will be cut into the following 4-gramms, if the word boundaries are respected and the words themselves are included: information info nfor form orma rmat mati atio tion, retrieval retr etri trie riev ieva eval.

[12] Petras/Perelman/Gey 2003. p. 243

Acknowledgments

For the calculation of data on the assessment Marco Winter gave a lot of support. The work on GIRT and in CLEF was partly funded by the European Commission under the IST program.

References

Alderuccio, D., Bordoni, L., Loretto, V. (2003): Data compression approach to monolingual GIRT task: an agnostic point of view. In Peters, C., Borri, F. (eds.) 2003), pp. 245-252

Basarnova, S., Magaj, H., Mdivani, R. Schott, H., Sucker, D. (eds.) (1997): Thesaurus Sozialwissenschaften Bd.1: Deutsch-Englisch-Russisch, Bd. 2: Russisch-Deutsch-Englisch, Bd. 3: Register. Bonn/Moskau: Informationszentrum Sozialwissenschaften / Institut für wissenschaftliche Information in den Gesellschaftswissenschaften (INION RadW)

Kamps, J., Monz, C., de Rijke, M., Sigurbjörnsson, B. (2003): The University of Amsterdam at CLEF 2003. In Peters, C., Borri, F. (eds.) (2003), pp. 71-78

Kluck, M. (1998): German Indexing and Retrieval Test Data Base (GIRT): Some Results of the Pre-test. In: Dunlop, Mark D. (Hrsg.): The 20th BCS IRSG Colloquium: Discovering New Worlds of IR (IRSG-98), Grenoble, France, 25-27 March 1998, Grenoble 1998 (= electronic workshops in computing) available: http://www.ewic.org.uk/ewic/workshop/view.cfm/IRSG-98

Kluck, M., Gey, F. C. (2001): The Domain-Specific Task of CLEF - Specific Evaluation Strategies in Cross-Language Information Retrieval. In Peters, C. (ed.) (2001), pp. 48-56

Kluck, M., Womser-Hacker, C. (2002): Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. In Rodríguez, M. G., Araujo, C. P. S. (eds.): Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas de Gran Canaria 29-31 May 2002 (pp. 573-576). Paris: ELRA

Kluck, M. (2002): Das Cross-Language Evaluation Forum (CLEF) - Evaluationsumgebung und Forschungskontext für mehrsprachiges Information Retrieval (mit einer Skizze der Ergebnisse von CLEF 2002). In Hammwöhner, R., Wolff, C., Womser-Hacker, C. (eds.): Information und Mobilität. Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002) (pp. 225-237). Konstanz: UVK

Kluck, M. (2003): Die Evaluation von Cross-Language-Retrieval-Systemen mit Hilfe der GIRT-Daten des IZ. Ein Bericht über die Entwicklung im Zeitraum von 1997 bis 2003. Bonn: Informationszentrum Sozialwissenschaften

Kluck, M. (2004): The GIRT Data as Means for Evaluation of CLIR-Systems – from 1997 until 2003. In Peters, C., Gonzales, J., Braschler, M., Kluck, M. (eds.) (2004): Proceedings of the Forth Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21 - 22, 2003 ; Revised Papers. Forthcoming in Lecture Notes in Computer Science

Leveling, J. (2003): University of Hagen at CLEF 2003: Natural Language Access to the GIRT4 Data. In Peters, C. Borri, F. (eds.) (2003), pp. 253-262

Peters, C. (ed.) (2001): Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Information Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers. Berlin et al.: Springer (Lecture Notes in Computer Science, 2069)

Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) (2002): Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001. Revised Papers. Berlin et al.: Springer (Lecture Notes in Computer Science, 2406)

Peters, C., Gonzales, J., Braschler, M., Kluck, M. (eds.) (2003): Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 19 - 20, 2002 ; Revised Papers. Berlin et al.: Springer (Lecture Notes in Computer Science; 2785)

Peters, C., Borri, F. (eds.) (2003): Results of the CLEF 2003 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway. Vol. I: Papers. Pisa: Centromedia

Peters, C., Braschler, M., Choukri, K., Gonzalo, J., Kluck, M. (2004): The Future of Evaluation for Cross-Language Information Retrieval Systems or CLEF: What Happens Now? (in this volume)

Schott, H. (ed.) (1999): Thesaurus Sozialwissenschaften – Thesaurus for the Social Sciences [Ausgabe – Edition] 1999. [Bd. 1:] Deutsch-Englisch – German-English, [Bd. 2] Englisch-Deutsch – English-German. Bonn: Informationszentrum Sozialwissenschaften

Vorhees, E. M., Harman D. K. (2000): Overview of the Eighth Text Retrieval Conference (TREC-8), 1999. In Vorhees, E. M., Harman D. K. (eds.): The Eighth Text Retrieval Conference (TREC-8). Gaithersburg: NIST, (pp. 1-23)

Petras, V., Perelman, N., Gey, F. C. (2003): UC Berkeley at CLEF 2003 – Russian Language Experiments and Domain-Specific Cross-Language Retrieval. In Peters, C. Borri, F. (eds.) (2003), pp. 235-244