Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment

Michael Kluck*, Christa Womser-Hacker[†]

*Social Science Information Centre (IZ)
Lennéstr. 30, 53113 Bonn, Germany
kluck@bonn.iz-soz.de
[†]University of Hildesheim, Institute for Applied Linguistics, Information Science
Marienburger Platz 22, 31141 Hildesheim, Germany
womser@uni-hildesheim.de

Abstract

Topic creation and relevance assessment are considered as crucial components of the evaluation process in Information Retrieval (IR). In the context of the Cross-Language Evaluation Forum (CLEF), the focus lies on evaluating multilingual functions of IR systems. Therefore, topics are generated in various languages and judging the documents delivered by the systems, requires native speakers of the participating languages who are experts in the topics' domains. In this paper, the important issues of topic generation and relevance assessment under multilingual conditions are discussed.

1. Introduction

The evaluation campaign of CLEF[1] comprises several components: the evaluation methodology, the evaluation software packages, the data collections, the topics, the overall results of the participants, the assessed results of the participants, and the calculated statistical results. In our paper we want to emphasize two of these components: the topics, and the assessments. In particular, we want to discuss the procedures and issues involved in these elements of the evaluation process under the special condition of multilinguality.

2. Topic creation at CLEF

Since the beginning of information retrieval evaluation, topic generation is considered as one of the most important tasks. In the course of the discussions on well-formed collections, which came up in the early 90s, simple queries changed to "original user requests" or so-called topics. Concerning the generation process, we could rely on the valuable experiences made in TREC[2] (Harman et al., 2001 and Voorhees, 2001) where some effects of topic characteristics have been investigated and changed over time. At the beginning, topics were very elaborated and carefully formulated so that the systems could start from a very good basis without applying any sophisticated query expansion techniques. Since this procedure seemed not to be very realistic, topics were formulated in a less structured and much shorter way (Sparck Jones, 2000).

Within the CLEF initiative, multilingual information retrieval is considered as main task. In 2001, five core languages (English, French, German, Italian, and Spanish) built the basis for documents and topics. There are some specific challenges of generating topics in a multilingual environment. To give same chances to each language, five language teams of the evaluation forum generated a certain number of topics in each of the core languages[3]. Participants could choose which languages should be their starting points for performing retrieval.

2.1. The topic creation process

2.1.1. Rules for creating topics

First of all, the CLEF language teams agreed upon a set of rules to establish common opinions on the topic generating process. The main goal was to create topics of real life which should meet the content of the documents ,which were drawn from different journals and newspapers of the year 1994. Political, social, cultural, economic, scientific, and sporting events were included. A specific structure similar to SGML with three textual fields was applied to the topics (see example below). The title field should sketch in a very short way the main content of the topic, the description field presents a more precise formulation in one sentence, and the narrative field states additional criteria concerning relevance. In accordance of the five teams, there should be so-called open topics and topics addressing specific facts. Roughly 20% should have answer documents containing fact information e.g. proper names or dates. Another heuristic rule said that topics should be related to either international, European or national events. It was not easy to satisfy this rule because of the very different scope of the newspapers. Local events which took place in the South of Germany e.g., were not reported upon in detail in the American Los Angeles Times or the Italian La Stampa.

The following example points out a typical CLEF topic:

[1] www.clef-campaign.org

[2] http://trec.nist.gov

[3] Further languages were supplemented (e.g. Dutch and Japanese).

```
<top>
<num>C088</num>
<EN-title>Mad Cow in Europe</EN-title>
<EN-desc>Find documents that cite cases of Bovine
Spongiform Encephalopathy (the mad cow disease) in
Europe.</EN-desc>
<EN-narr>Relevant documents will report statistics and/or
figures on cases of animals infected with Bovine
Spongiform Encephalopathy (BSE), commonly known as
the mad cow disease, in Europe. Documents that only
discuss the possible transmission of the disease to humans
are not considered relevant.</EN-narr>
</top>
```

Figure 1. Example of a topic

2.1.2 Invention and proposal of topics

Each language team generated a set of 15 possible topics, which was much more than needed. To be sure of the existence of related documents, pre-search processes were performed in the databases of the five languages. During a meeting of all language teams all topic suggestions were discussed intensively with respect to their content and their formulation. A set of 50 topics was selected with origins in the different languages. The groups elaborated this basic set and translated it from their original languages to English. The goal was to receive reliable formulations in the five languages. This process was performed in a very communicative and cooperative way.

2.1.3. Translation of the topic set

Translation processes aim at transferring an original text from one language to a second target language. As we know from translation science, the linguistic and the cultural background are very important. Very rarely, a simple one-by-one translation is possible. In a multilingual environment, by translating the topics different problems arise which should be met by the systems' functionalities. The most important of these challenges deal with proper names, abbreviations, compounds, idiomatic etc. Domain specific terminology and culture specific knowledge are involved.

2.1.4. Cross-checking of the topic translations

The quality of the final topic set (i.e. 50 topics in English, French, German, Italian and Spanish) was checked by professional translators. The most important modifications referred to stylistic, grammatical, semantic categories, but also the correction of typos and formal mistakes was proposed. The main issue was that the translated topics in all aspects exactly referred to the same content. All suggested modifications were discussed with the topic generators to ensure to maintain the intended meaning.

The following examples show that simple word-by-word translations are not adequate in this case. Often, additional explanations or the resolution of culture-specific terms or abbreviations were necessary:

ORIGINAL LANGUAGE	TARGET LANGUAGE
EN “CNG cars”	DE “mit Flüssiggas betriebene Autos”
DE “Schneider-Konkurs”	FR „Faillite de M. Schneider”
NL “Muisarm”	FR “ordinateur: souris et tensions musculaire”
ES “Subasta de objetos de Lennon”	FR “Vente aux enchères de souvenir de John Lennon”
DE “deutsche Spätaussiedler”	EN “people of German origin from Eastern Europe coming to live in Germany”

Table 1: Examples of translation problems

2.2. Issues of making topics challenging

In CLEF, topics were not constructed artificially but formulated in a natural way. I.e., that linguistic properties are distributed by chance. The following table shows the differences between English and German.

	ENGLISH	GERMAN
Stemming	345	471
Compound words	9	115
Proper Names	73	62
Abbreviations	14	13
Negations	18	21
Idioms	2	2
Dates	12	12
Noun Phrases	98	42

Table 2: Analysis of English and German Topics in CLEF 2001

Table 1 reflects the various language characteristics. In the German language e.g. more linguistical problem arose concerning stemming and decomposition of words. On the other hand, the English topics contained more noun phrases than the German ones. Facts were distributed more or less equally over the two languages.

3. Relevance assessment

3.1. Applying the TREC methodology

The CLEF relevance assessment process is based on the methodology and experiences from the TREC campaigns (Voorhees and Harman, 2001). The general evaluation procedures of TREC have also been used for CLEF. This means mainly the pooling method, which is applied to large test collections as they are also used in CLEF. The pooling method creates a subset of documents out of the whole collection to be judged for a specific topic. In the CLEF campaign of 2001 for each run[4] included in the pool, the top 60 documents per topic have been added to the pool of this topic. Those 60 documents are seen to be most likely relevant to that topic, since the retrieved results are delivered by the participating systems in a ranked list with decreasing order of relevance. For a

[4] Run means each set of results for all topics which has been treated with a different retrieval methodology and/or a different retrieval software and has been delivered by a participating group within a specific track of CLEF.

given topic many documents are retrieved for more than one run of a system or even for more than one system, so the pool is smaller than the possible maximum (in case of CLEF about a quarter of the maximum size). Unjudged documents (those not included in the pool) are assumed to be not relevant.

The process of the assessment itself is based on a binary judgement of the respective assessor whether a given document is relevant with respect to a specific topic or not. The assessors shall assume that they are writing a report on the given topic, and they should include any document that contains relevant information on this topic (where the document is relevant as a whole or partially). They should judge a document as relevant regardless of other documents even if they are containing the same information.

3.2. Relevance assessment as part of comparative evaluation

Research on the evaluation procedures of TREC has shown that this methodology is appropriate to such a comparative evaluation scenario (Voorhees, 2000 and 2001; Zobel 1998). The goal of the TREC (and CLEF) evaluation method is to compare the outcome of retrieval systems, that means, the result figures give relative scores of evaluation measures, not absolute ones. The differences in relevance judgements, which really occur, do not matter as far as the relative measures based on these judgements do not change significantly. The variations of judgements by different assessors and for the same assessor over time do not affect the comparative evaluation. On the other hand the assumption is that a sufficient number of included runs will turn up the most of the relevant documents. If a system did not contribute to the pool of judged documents, it might be unfairly penalised by the evaluation statistics. But both effects have also been investigated for CLEF by Braschler (2001) at the CLEF 2000 Workshop and Braschler (2002) and Hiemstra (2001) at the CLEF 2001 Workshop, and they have proved that the evaluation measures are reliable, stable, and fair to all tested systems.

3.3. Specific problems of multilingual assessment

Doing Cross-Language Information Retrieval (CLIR) means to have additional problems with assessments to be solved like obtaining as much as possible consistency of the relevance assessment of the topics per language and among languages. The consistency of judgements is much harder to obtain, because there are multiple assessors per topic (one per language group). There are some measures taken to work as consistent as possible:

1. to include the assessors (or at least the coordinators of the assessments in the different language groups) in the topic creation phase and in the discussion as well on the definite wording of the topics as on the selection of the definite topic set.

2. to cross-check the translations of the topics into the different languages to avoid hidden changes in the meaning. Nevertheless sometimes it is not possible to have a direct translation because the respective concept does not exist in this language (see table 1); then the translation gives a more vague description of the concept that is very clear in the other language. Here the assessors have the possibility to look at the original topic in the original language (or other languages) to get some clarification of the meaning (if they are able to understand this language).

3. to communicate immediately on occurring problems during the assessment phase.

4. to try a two-stage approach: in a first run decide on the clearly relevant documents; in a second run discuss the unclear cases with the supervisor or another assessor of the same language group.

3.4. General problems evolving during the assessment

Although in this context the organisation of the topic generation process plays an important role for assuring as much as possible consistency among the language groups doing the assessments, additional problems evolve.

The procedure doing the assessment is as follows: First of all the assessors should read the whole topic carefully. They should read all parts (title, description, narrative) and especially take the narrative as the definition of what is meant to be relevant (or not). In some cases (when the topic is asking for a factual answer like a proper name or a date etc.) the assessors have the correct information at hand that is given as an addendum to the narrative (which is of course not shown to the participants). Additionally, the assessment software that has been provided by NIST allows predefining relevant terms or words which probably express the information need that lies behind the topic text. These terms are highlighted by the system whenever they occur in the documents to be assessed. Thus, in most cases the important sentences or phrases within the documents are already indicated by the highlighting functionality. As a document counts as relevant if at least a part of it contains relevant information, the assessors have a good support to find the relevant parts of the document easily. On average, there are between 200 and 500 documents to be judged for one topic in each language in CLEF.

It is possible to start or stop the judgement process at any time because the system saves all actions and judgements, thus, you can restart as often as you want or need. But it is more convenient and leads to more concise results if an assessor goes through a topic within one step and without a restart, because she or he may change the criteria implicitly in between.

On the other hand, there is a tendency of shifting the assessment over the time of assessing documents on the same topic. This danger occurs mostly if for a given topic no or very few documents are relevant. In this case the assessors tend to become less restrictive with their judgement the more documents they judge. But also the opposite tendency is occurring: if there are too many documents that seem to be relevant, the criteria of relevance might be tightened over time. Against both tendencies the only way is that the assessors reassure themselves from time to time during the assessment process of a topic what the topic (which means the information need expressed by the topic) really is: they have to re-read the whole topic carefully.

Another problem occurs when the assessors discover during the assessment of a topic that he or she has overlooked certain aspects of the topic or certain other wordings for the topic or for its aspects. This may also be caused by different wording in different regions of a language area (like German in Germany or German in Switzerland, French in France or in Canada, American English or British English): for example “Abschiebung” is used in Germany and “Ausschaffung” is used in Switzerland for the same fact: here the deportation or extradition. If this case of overlooking aspects or wordings occurs, all recent judgements have to be re-assessed in order to include the additional aspect or the additional wordings into the assessment.

Sometimes it occurs during the assessment of a specific topic that the topic becomes unclear or new aspects show up: this may arise doubt about the real meaning of the topic or its range. This happens even if the topic seemed to be very clear at the first glance (and also during the topic creation meeting of the language groups). In this case a short discussion by e-mail between the language groups helps to clarify the meaning and (much more important) to assure the common understanding of relevance.

4. Conclusions and further research questions

Topic generation and relevance assessment are very important components of IR evaluation methodology. Further analyses are necessary to get more experience in formalising topic characteristics and investigating what makes a topic difficult or challenging for IR system evaluation. Concerning the assessment process there are also open research questions, and the growing set of results and experiences will give the background for future research.

To assure the reliability of the pools it should be investigated how the inclusion of runs that not already have been included into the pool would effect the pool and the comparative results. This would mean to add further relevance assessments to the assessment phase or in an additional assessment phase.

Another interesting topic would be to compare the relevant documents found during the topic creation process and those retrieved by the participants’ systems. This could give some insights why some systems did not find the documents which have been found by humans by intellectually formulating queries with ZPRISE which is used by most of the language groups for testing topics before designing the definite topic set.

To reassure the validity of the assessments the result sets could be assessed by a second assessor for some of the topics and the precision-recall figures could be computed with the different relevance judgements and their union.

Some considerations and tests should be carried out to stabilise the number of judged documents per language. For instance more monolingual runs could contribute (as sort of baselines) to the pool.

5. References

Braschler, M., 2002 CLEF 2001 - Overview of Results. In M. Braschler, J. Gonzalo, M. Kluck, C. Peters (eds.), *Cross Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 2001, Revised Papers*. Berlin et al.: Springer (in print)

Braschler, M., 2001 CLEF 2000 - Overview of Results. In C. Peters (ed.), *Cross Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 2000, Revised Papers*. Berlin et al.: Springer (LNCS 2069), 89-101

Harman, D., M. Braschler, M. Hess, M. Kluck, C. Peters, P. Schäuble, and P. Sheridan, 2001. CLIR Evaluation at TREC. In C. Peters (ed.), *Cross Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 2000, Revised Papers*. Berlin et al.: Springer (LNCS 2069), 7-23

Hiemstra, D. 2001. The CLEF Relevance Assessment in Practice. *Talk at the CLEF 2001 Workshop, 3 September, Darmstadt, Germany*

Kluck, M., T. Mandl and C. Womser-Hacker 2002. Cross-Language Evaluation Forum (CLEF) – Europäische Initiative zur Bewertung sprachübergreifender Retrievalverfahren. In *Information Wissenschaft & Praxis* 53:82-89

Sparck Jones, K., 2000. Further Reflections on TREC. *Information Processing & Management* 36:37-88.

Voorhees, E., 2001. Philosophy of IR Evaluation. In C. Peters (ed.): *Results of the CLEF 2001 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2001 Workshop, 3 September, Darmstadt, Germany*, Sophia-Antipolis: ERCIM, 257-260

Voorhees, E., 2000. Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness. *Information Processing & Management* 36:679-716

Voorhees, E. and D. Harman, 2001. Overview of the Ninth Text Retrieval Conference (TREC-9). In E. Voorhees and D. Harman (eds.) *The Ninth Text REtrieval Conference (TREC 9)*. Gaithersburg: NIST, 1-14,http://trec.nist.gov/pubs/trec9/t9_proceedings.html

Zobel, J., 1998. How Reliable Are the Results of Large-Scale Information Retrieval Experiments? In W. Bruce, A. Moffat, C.J. van Rijsbergen, R. Wilkinson and J. Zobel (eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, 307-314