Manual Queries and Machine Translation in Cross-language Retrieval and Interactive Retrieval with Cheshire II at TREC-7

Fredric C. Gey, Hailing Jiang and Aitao Chen
UC Data Archive & Technical Assistance (UC DATA)
gey@ucdata.berkeley.edu, hjiang1@sims.berkeley.edu, aitao@sims.berkeley.edu

Ray R. Larson
School of Information Management and Systems
ray@sherlock.berkeley.edu
University of California at Berkeley, CA 94720

Abstract

For TREC-7, the Berkeley ad-hoc experiments explored more phrase discovery in topics and documents. We utilized Boolean retrieval combined with probabilistic ranking for 17 topics in ad-hoc manual entry. Our cross-language experiments tested 3 different widely available machine translation software packages. For language pairs (e.g. German to French) for which no direct machine translation was available we made use of English as a universal intermediate language. For CLIR we also manually reformulated the English topics before doing machine translation, and this elicited a significant performance increase for both quad language retrieval and for English against English and French documents. In our Interactive Track entry eight searchers conducted eight searches each, half on the Cheshire II system and the other half on the Zprise system, for a total of 64 searches. Questionnaires were administered to gather information about basic demographic and searching experience, about each search, about each of the systems, and finally, about the user's perceptions of the systems.

1 Introduction

Berkeley's participation in the TREC conferences has been used as a testing ground for algorithms for probabilistic document retrieval. Probabilistic document retrieval attempts to place the ranking of documents in response to a user's information need (generally expressed as a textual description in natural language) on a sound theoretical basis. Bayesian inference is applied to develop predictive equations for probability relevance where training data is available from past queries and document collections. Berkeley's particular approach has been to use the technique of logistic regression. Logistic regression has by now become a standard technique in the discipline of epidemiology for discovering the degree to which causal factors result in disease incidence [9]. In document retrieval the problem is turned around, and one wishes to predict the incidence of a rare disease called 'relevance' given the evidence of occurrence of query words and their statistical attributes in documents.

In TREC-2 [3] Berkeley introduced a formula for ad-hoc retrieval which has produced consistently good retrieval results in TREC-2 and subsequent TREC conferences TREC-4 through

TREC-6. The logodds of relevance of document D to query Q is given by

$$\log O(R|D,Q) = -3.51 + \frac{1}{\sqrt{N}+1}\Phi + 0.0929 * N \tag{1}$$

$$\Phi = 37.4\sum_{i=1}^{N}\frac{qtf_i}{ql+35} + 0.330\sum_{i=1}^{N}\log\frac{dtf_i}{dl+80} - 0.1937\sum_{i=1}^{N}\log\frac{ctf_i}{cf} \tag{2}$$

where

N	is the number of terms common to query and document,
qtf_i	is the occurrence frequency within a query of the ith match term,
dtf_i	is the occurrence frequency within a document of the ith match term,
ctf_i	is the occurrence frequency in a collection of the ith match term,
ql	is query length (number of terms in a query),
dl	is document length (number of terms in a document), and
cf	is collection length, i.e. the number of occurrences of all terms in a test collection.

The summation in equation (2) is carried out over all the terms common to query and document. This formula has also been used, with success, in document retrieval with Chinese and Spanish queries and document collections of the past few TREC conferences. In TREC-6, we utilized this identical formula for German queries against German documents in the cross-language track for TREC-6. In TREC-7 this was the formula also used for all cross-language runs.

2 Ad-hoc retrieval

In TREC-6 [7], Berkeley introduced a variation of the formula which explicitly separated the evidence supported by phrases from the evidence supported by single terms. Phrases were chosen using a technique from computational linguistics, computation of the Mutual Information (MI) Measure which showed whether two words occurred together more than randomly. However, in the Berkeley TREC-6 experiments and subsequent experiments showed no discernible advantage to separability of phrases.

For TREC-7 Berkeley experimented with different stemming and phrase discovery approaches which fall short of full NLP tagging of phrases, including modification of the MI measure to be used after stemming and use of the WordNet stemmer. For example with TREC-7 topic 379 “mainstreaming” the Lovins-style stemmer of SMART-11 truncates to the fairly common term “mainstream,” while the WordNet stemmer leaves the term as a whole. This term was used in our manual submission and the resulting precision for that query moved from 0.0244 for the fully automatic run (using the SMART-11 stemmer) to 0.3658 using a Boolean query (described below) and the WordNet stemmer. In the large, however, our experiments showed no significant advantage of one stemmer over another.

One failure of phrase discovery directly derived from our decision to abandon phrase discovery before stop word processing (which we had done in TREC-6). Phrase discovery before stop word processing required us to maintain a file of a large number of bigrams (word pairs) which was too big for our system to maintain efficiently. However, for the topic 368 “in-vitro fertilization” the phrase “in-vitro” can’t be found because “in” is a stop word. Another failure of phrase discovery occurred in topic 394 “home schooling” – the words ‘home’ and ‘school’ are very common and hence an MI measure does not discover this term, whereas Natural Language Processing of this topic would surely uncover this crucial term. Our best performance on this query had overall precision 0.0538.

2.1 Boolean queries

Past research by Hearst [8] and Cormack et al. at Waterloo [6] has indicated that a carefully constructed Boolean query can be used to weed out irrelevant documents and thereby increase the precision of other selected documents. Berkeley decided to experiment with this approach in a limited way. Seventeen topics were given a Boolean formulation (actually many more were experimented with, but for these seventeen it seemed that an improvement might be obtained over automatic full text and manual queries.

In almost all cases (14 of the 17 topics), an improvement resulted, but a spectacular improvement occurred for three topics. First, for topic 351 "Falkland petroleum exploration" our title and narrative runs (Brkly24, Brkly25) had precision 0.2982 and 0.3137, whereas the Boolean query (in prefix form)

> (AND (OR falklands falkland_islands) (OR Britain UK Argentina) (OR oil petroleum))

achieved a precision of 0.8784, best TREC-7 overall run for that query. For topic 352 "British Chunnel Impact" the Brkly 24 and Brkly25 runs were an abysmal 0.0379 and 0.0097 respectively while the Boolean query

> (AND (OR British Britain English) (OR chunnel (AND channel tunnel)))

obtained a precision of 0.3112. Finally for the above mentioned topic 379 "Mainstreaming" we formulated the following Boolean query:

> (OR (AND mainstreaming education) (AND mainstream schools) (AND handicapped schools))

to obtain the precision of 0.3658.

Boolean queries, of course, return an unranked set of documents, almost always fewer than the 1,000 documents required by TREC for ranking systems. So how should one rank the set of documents retrieved by a Boolean query, and how should one augment a retrieved set size less than 1,000 documents? Berkeley's approach was to use its standard logistic regression ranking algorithm for the Boolean query's document set, and to make a separate run of all manually reformulated queries ranked using logistic regression, and then to merge the two by adding the value of 1 to all documents in Boolean set retrieved. Then we have the problem of the same document appearing twice in the ranked set with different estimated probability of relevance. This was resolved by removing duplicates from the Boolean retrieved set before ranking it. An alternative would be to remove the lower ranked duplicate, but we choose not to do this.

3 Cross-language Retrieval Experiments

We created one index file from TREC-7 CLIR collections consisting of documents in English, French, German and Italian. The English words are stemmed but not the French, German and Italian words. For English, we used the SMART stemmer and a list of some 600 stopwords. We constructed a French stopword list by combining the French translation of the English stopwords using SYSTRAN [13] and the top 200 French words that most frequently occur in the French document collection. The German stopword list and the Italian stopword list were constructed in the same manner.

We submitted four cross-language retrieval runs using queries in English, French, German, and Italian against documents in all four languages. Our approach to the CLIR task is to translate the queries in the source language to other languages that are present in the collection using machine translation software. The copy of the Globalink [2] we used is capable of translating English to French, German and Italian and vice versa; however, the translation among French, German adn Italian is not supported. For the English queries, we directly translated them into French, German and Italian using the Gloablink machine translation software. But for the French, German and Italian queries, we had to use Enlgish as a universal intermediate language. For example, the French queries were translated into English using Globalink; then the English translations of the original French queries were translated into German and Italian using Globalink again. The process of translating French queries into English, German and Italian is illustrated in Figure 2. For each set of queries in a source language, we have generated three set of queries in the other three query languages. For each set of queries, the translations and the source queries were combined to produce a set of multilingual queries. The pooled multilingual queries were run against the document collection consisting of documents in four languages. The final results for each run consists of the top-ranked 1000 documents for each pooled query. The translation and retrieval process of using English queries as the source queries is illustrated in Figure 1. The results of our four official runs are presented in Table 1. For the BKYCL7ME run, the English queries were manually reformulated before they were translated into other languages.

Run ID	Category	Query Language	Document Languages	Average Precision	Relevant Retrieved	No. >= Median	No. < Median
BKYCL7ME	Manual	English	E,F,G,I	0.3390	2648	23	5
BKYCL7AF	Automatic	French	E,F,G,I	0.2369	2405	12	16
BKYCL7AG	Automatic	German	E,F,G,I	0.2406	2482	12	16
BKYCL7AI	Automatic	Italian	E,F,G,I	0.2184	2344	12	16

Table 1: Results of four official runs.

After the relevance judgments for the cross-language retrieval were made available, we performed two additional runs using English queries against French and English documents. The results for those two runs are shown in Table 2. Our manual run of English queries against English and French documents performed substantially better than the automatic run. We also evaluated three machine translation software packages—SYSTRAN, Globalink, and EasyTranslator [1]—on the TREC-7 CLIR test collection. The average precision values over a set of 28 queries are shown in Table 3

Run ID	Category	Query Language	Document Languages	Average Precision	Relevant Retrieved	No. >= Median	No. < Median
BKYCL7MEF	Manual	English	E,F	0.4185	2106	27	1
BKYCL7AEF	Automatic	English	E,F	0.3261	2007	23	5

Table 2: Results of English queries against English and French collections.

The Globalink translations show that Globalink leaves new words (i.e., words unknown to the translation system) in the source text unchanged. The term mismatch problem arises when the spellings of the equivalents of a word are different and the word is left untranslated. For example, in topic 26, the German equivalent of the proper name *Létschberg* in English is *Lötschberg* and the Italian equivalent is *Lütschberg*; the French equivalent is the same as the English one. Because the same proper name has different spellings in English, German and Italian, we believe that the failure of properly translating the proper name in one language into its equivalents in other

	English (Manual)	English	French	German	Italian
SYSTRAN	0.3316	0.2615	0.2318	0.2102	0.1924
Globalink	0.3390	0.2602	0.2369	0.2406	0.2184
EasyTranslator	0.3072	0.2302	0.1795	0.1961	

Table 3: Comparison of three machine translation systems in cross-language retrieval.

languages would result in missing many of the relevant documents in the multilingual collection.

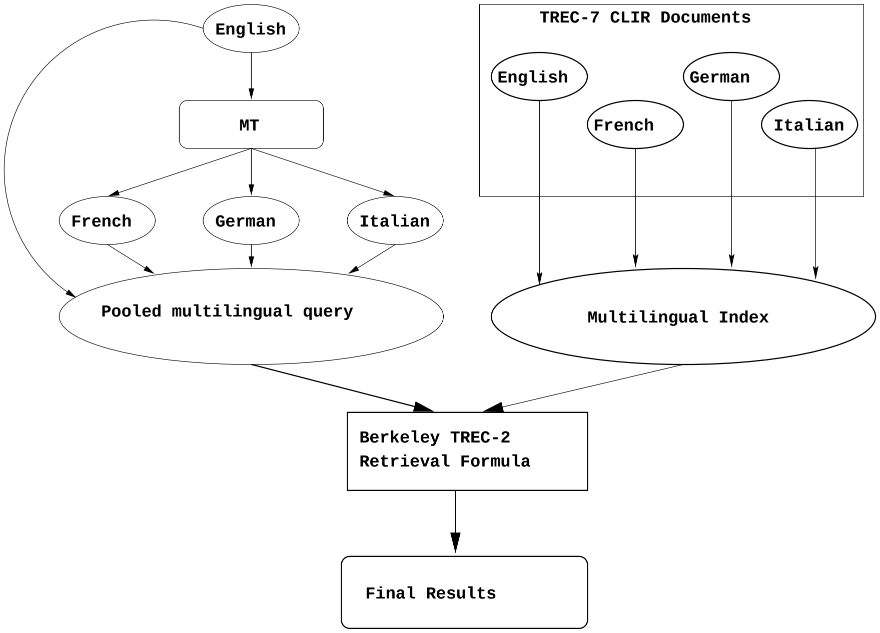


Figure 1: TREC-7 CLIR.

4 Interactive Probabilistic Retrieval: Cheshire II at TREC 7

This section briefly discusses the UC Berkeley entry in the TREC7 Interactive Track. In this year's study eight searchers conducted eight searches each, half on the Cheshire II system[11] and the other half on the Zprise system, for a total of 64 searches. Questionnaires were administered to gather information about basic demographic and searching experience, about each search, about each of the systems, and finally, about the user's perceptions of the systems. This section will briefly describe the systems used in the study and how they differ in design goals and implementation. The results of the interactive track evaluations and the information derived from the questionnaires are then discussed and future improvements to the Cheshire II system are considered. A more detailed version of the discussion in this section is available as http://sherlock.berkeley.edu/cheshire_trec7.pdf.

The primary goals of UC Berkeley entry in the TREC-7 Interactive track were to 1) attempt to replicate our entry in the TREC-6 Interactive track[10] with a larger number of participants (searchers), and to see if there were substantial differences in the ranking of the systems between last year and this year, and 2) to follow the complete TREC-7 Interactive track protocol to obtain further information than obtained in TREC-7 via the standard questionnaires filled in by all searchers on all systems. We are hoping to develop a baseline that can be used to evaluate changes and additions

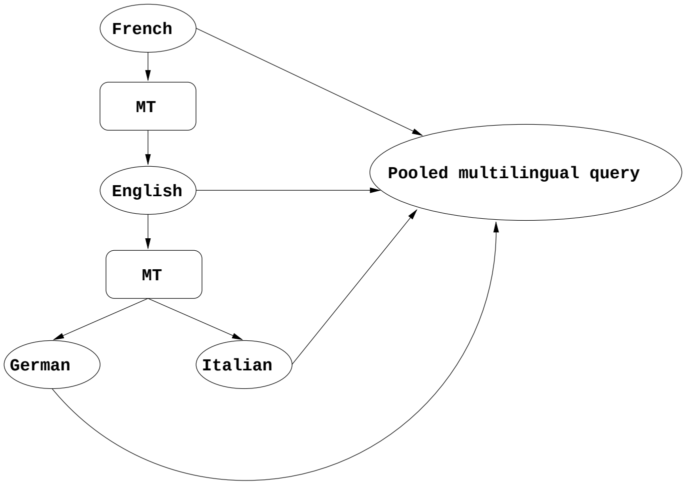


Figure 2: Query Generation.

to the systems (primarily to the Cheshire II system) in the future.

In TREC-7 we used virtually identical implementations of the Cheshire II system and the ZPRISE system as those used in TREC-6. The database and indexing for each system were also the same as for TREC-6. The characteristics of the Cheshire II system and ZPRISE systems are discussed below.

4.1 The Cheshire II System

The original design goals of the Cheshire II system were to develop a "next-generation" online library catalog system that would incorporate ranked retrieval based on probabilistic retrieval methods along with the Boolean retrieval expected in "second generation" online catalog systems. Much has changed since these initial goals were formulated. The Cheshire II system now finds its primary usage in full text or structured metadata collections based on SGML and XML, often as the search engine behind a variety of WWW-based "search pages" or as a Z39.50 server for particular applications.

4.1.1 The Cheshire II Search Engine

The Cheshire II search engine supports both probabilistic and Boolean searching. The design rationale and features of the Cheshire II search engine have been discussed in the TREC-6 paper [10] and will only be briefly repeated here.

The search engine functions as a Z39.50 information retrieval protocol server providing access to a set of databases. In the TREC-7 experiments the TREC Financial Times (FT) database was the only database used by participants. The system supports various methods for translating a searcher's query into the terms used in indexing the database. These methods include elimination of unused words using field-specific stopword lists, particular field-specific query-to-key conversion or "normalization" functions, standard stemming algorithms (Porter stemmer) and support for mapping database and query text words to single forms based on the WordNet dictionary and thesaurus using a adaption of the WordNet "Morphing" algorithm and exception dictionary..

The Cheshire II search engine supports both Boolean and probabilistic searching on any indexed element of the database. In probabilistic searching, a natural language query can be used to retrieve the records that are estimated to have the highest probability of being relevant given the user's query. The search engine supports a simple form of relevance feedback, where any items found in an initial search (Boolean or probabilistic) can be selected and used as queries in a relevance feedback search.

The probabilistic retrieval algorithm used in the Cheshire II search engine is based on the *logistical regression* algorithms developed by Berkeley researchers and shown to provide excellent full-text retrieval performance in previous TREC evaluations [5, 3, 4]. Formally, the probability of relevance given a particular query and a particular record in the database $P(R \mid Q, D)$ is calculated and the records are presented to the user ranked in order of decreasing values of that probability. In the Cheshire II system $P(R \mid Q, D)$ is calculated as the "log odds" of relevance $\log O(R \mid Q, D)$, where for any events A and B the odds $O(A \mid B)$ is a simple transformation of the probabilities $\frac{P(A|B)}{P(\bar{A}|B)}$. The Logistic Regression method provides estimates for a set of coefficients, c_i, associated with a set of S statistics, X_i, derived from the query and database, such that

$$\log O(R \mid Q, D) \approx c_0 \sum_{i=1}^{S} c_i X_i \tag{3}$$

where c_0 is the intercept term of the regression.

For the set of M *terms* (i.e., words, stems or phrases) that occur in both a particular query and a given document.

The regression equation and coefficients used in the TREC-7 Interactive Track are the same as used in our TREC-6 entry. These are based on the TREC-3 Adhoc entry from Berkeley [4] where the coefficients were estimated using relevance judgements from the TIPSTER test collection:

The basic elements are:

$X_1 = \frac{1}{M} \sum_{j=1}^{M} logQAF_{t_j}$. This is the log of the absolute frequency of occurrence for term t_j in the query averaged over the M terms in common between the query and the document. The coefficient c_1 used in the current version of the Cheshire II system is 1.269.

$X_2 = \sqrt{QL}$. This is square root of the query length (i.e., the number of terms in the query disregarding stopwords). The c_2 coefficient used is -0.310.

$X_3 = \frac{1}{M} \sum_{j=1}^{M} logDAF_{t_j}$. This is is the log of the absolute frequency of occurrence for term t_j in the document averaged over the M common terms. The c_3 coefficient used is 0.679.

$X_4 = \sqrt{DL}$. This is square root of the document length. In Cheshire II the raw size of the document in bytes is used for the document length. The c_4 coefficient used is -0.0674.

$X_5 = \frac{1}{M} \sum_{j=1}^{M} logIDF_{t_j}$. This is is the log of the *inverse document frequency*(IDF) for term t_j in the document averaged over the M common terms. IDF is calculated as the total number of documents in the database, divided by the number of documents that contain term t_j The c_5 coefficient used is 0.223.

$X_6 = logM$. This is the log of the number of common terms. The c_6 coefficient used in Cheshire II is 2.01.

These coefficients and elements of the ranking algorithm have proven to be quite robust and useful across a broad range of document types.

Probabilistic searching, as noted above, requires only a natural language statement of the searcher's topic, and thus no formal query language or Boolean logic is needed for such searches. However, the Cheshire II search engine also supports complete Boolean operations on indexed elements in the database, and supports searches that combine probabilistic and Boolean elements. At present, combined probabilistic and Boolean search results are evaluated using the assumption that the Boolean retrieved set has an estimated $P(R \mid Q_{bool}, D) = 1.0$ for each document in the set, and 0 for the rest of the collection. The final estimate for the probability of relevance used for ranking the results of a search combining Boolean and probabilistic strategies is simply:

$$P(R \mid Q, D) = P(R \mid Q_{bool}, D) P(R \mid Q_{prob}, D) \tag{4}$$

where $P(R \mid Q_{prob}, D)$ is the probability estimate from the probabilistic portion of the search, and $P(R \mid Q_{bool}, D)$ the estimate from the Boolean. This has the effect of restricting the results to those items that match the Boolean portion, with ordering based on the probabilistic portion.

Relevance feedback is implemented quite simply, as probabilistic retrieval based on extraction of content- bearing elements (such as titles, subject headings, etc.) from any items that have already been seen and selected by a user. Similarly, multiple records may be selected and submitted for feedback searching. In this case the contents of all those records are merged into a single query and submitted for searching. At the present time we do not use any methods for eliminating poor search terms from the selected records, nor special enhancements for terms common between multiple selected records [12], but we plan to experiment further with various enhancements to our relevance feedback method.

4.1.2 The Cheshire II Client Interface

The design of the Cheshire II client interface (shown with the TREC FT database in Figure 3) was driven by a number of goals:

1. to support a consistent interface to a wide variety of Z39.50 servers., and to dynamically adapt to the particular server.

2. to reduce the cognitive load on the users wishing to interact with multiple distributed information retrieval systems by providing a single interface for them all.

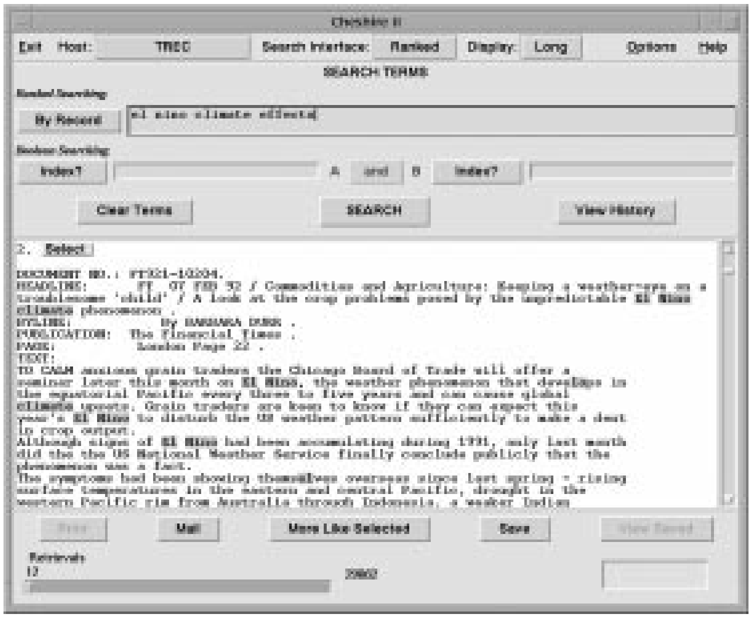
3. to minimize use of additional windows during users' interactions with the client in order to allow them to concentrate on formulating queries and evaluating the results, and not expend additional mental effort and time switching their focus of attention from the search interface to display clients;

4. to provide functions not immediately related to searching, such as print and e-mail facilities, to facilitate users' ability to 'take the results home'; and

5. to design a help system within the interface that would assist users not only in the mechanics of operating the Cheshire II client, but also in the more general tasks of selecting appropriate resources for searching, formulating appropriate queries, and employing various search tactics.

However, the initial design goals for the client interface made the assumption that most of the information that would be viewed in the search interface would be brief metadata records for documents, and not full text documents themselves. The ability to view full-text documents such as the FT articles used in the TREC-7 Interactive track experiments was added to the existing interface easily, but as the comments and questionnaire responses discussed below show, this was probably not an optimal implementation for the tasks posed by the experiments.

Additional functionality beyond searching and browsing has been relatively easy to implement. Functions for printing, e-mailing and saving records are all available when records are displayed, and the user has the option of acting on either the entirety of the current record display or a subset



el nino climate effects

Figure 3: Cheshire II Long Display.

thereof by selecting individual records using the "select" buttons on each record (visible in Figure 3 next to the record number).

Among the changes made to the client interface for TREC was the inclusion of display formats for the FT records (as shown in Figure 3). A routine was also added to highlight query terms in the text of the document to aid searchers in scanning for relevant passages. Note that the highlighting feature doesn't necessarily catch all of the terms that contributed to the selection of the document, because only the original query terms, and not stemmed terms, are used in the highlighting. Since the highlighting is using simple string matching on the text, partial words are sometimes highlighted erroneously.

4.2 The Zprise System

The second (control) system used in the TREC-7 Interactive track at Berkeley was the Zprise system from NIST. This system was used in the same configuration and with the same database indexing setup as used for the global control system in the TREC-6 Interactive Track. Zprise, as configured for this test was limited to a total of 24 retrieved items and relevance feedback was disabled. However, the interface was set up so that it provided a very good fit for the tasks involved in the interactive track. For example, documents were viewed in full text form in a separate window from the short display (consisting primarily of title and date as well as control elements for indicating relevant documents and for moving around in the brief display(see Figure 4.

Most of our users found the ZPRISE displays simple to learn and to operate, in fact most found that the operations required to carry out the Interactive Track tasks were easier to do on the ZPRISE interface than they were on the Cheshire II interface. This was not entirely surprising, since the ZPRISE interface is designed to support TREC-like databases containing full text, while the Cheshire II interface, as noted above, was designed for brief metadata records and not with the idea of providing support for the sort of reading and selection activities that make up the user tasks in the TREC Interactive Track.

In some ways the comparison between the interfaces comes down to how well a generic interface,

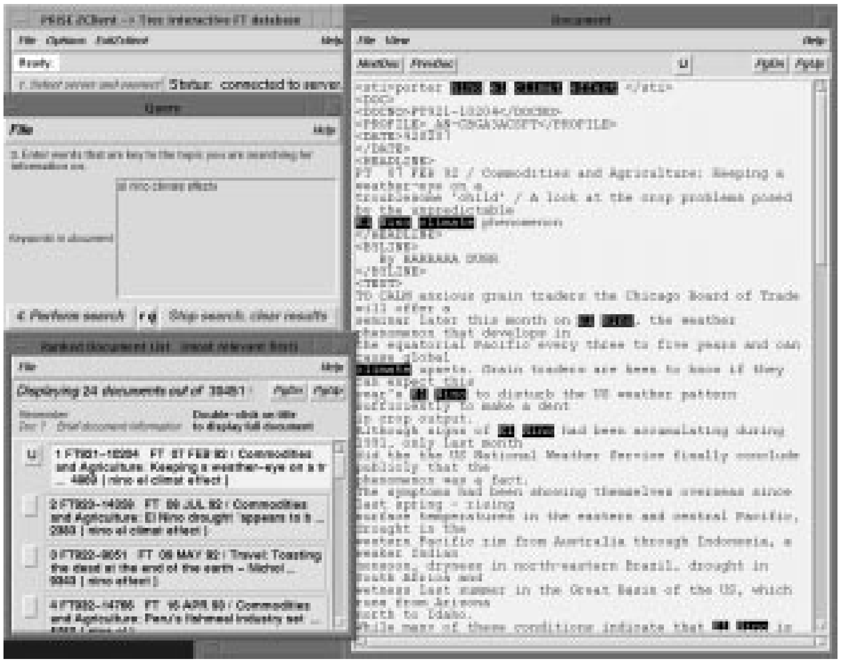


Figure 4: ZPrise Interface.

not particularly adapted to the specific task, compares to an interface tailored to that task. The underlying PRISE search engine in ZPRISE uses (apparently) a fairly standard Vector Space model search algorithm, which performs quite well given the usually brief and simple query statements that characterized most searches by the searchers participating in this year's interactive track. In the following section I will describe the results obtained from the Aspectual Precision and Recall evaluations at NIST and the results of the demographic, search and system related questionnaires filled out by the participating searchers.

4.3 TREC Interactive Track

The administration of the interactive track followed the protocols set down in the track guidelines. This mandated a minimum group of 8 participant searchers, each of whom conduct 8 searches, half on the control system (ZPRISE, identified as "Z") and half on the experimental system (Cheshire II, identified as "C").

Each searcher was asked to use the features of the respective interfaces to select as relevant those documents that they considered to relevant to one or more aspects of the specific topic. Because of some delays in obtaining the license and materials for the FA-1 Controlled Associations Test, this test was administered to the subjects independently from their actual search sessions.

The pooled results for all systems were evaluated at NIST by the TREC evaluators and "Aspectual Precision" and "Aspectual Recall" for each searcher was calculated. Table 4 shows the values for Aspectual Precision by TREC topic for all systems in the TREC Interactive Track. Table 5 shows the values for Aspectual Recall for all of the participating systems. Note that these two tables were derived from the per- search Recall and Precision figures reported by NIST. Note also that in these tables all of the system usages were combined in the calculations, therefore "ok_noRF" which was used in two separate experiments has the results from both experiments combined. The two Berkeley systems ("C" and "Z", the Cheshire II system and ZPRISE systems respectively) are shown in boldface in Tables 4 and 5. The control system "Z" performed marginally better than the experimental system in terms of the Aspectual Precision. It is also interesting to note that virtually identical performance was achieved by the "ZP" system from NMSU and the "zp_noRF" from the Okapi Group, I believe that both of these systems, like "Z", are unmodified ZPRISE systems.

The Cheshire II system also did not perform as well the control ZPRISE system in these experiments. This fact can largely be attributed to the more complex interactions required to perform the search tasks on the generic Cheshire II interface than on the ZPRISE system. In addition, there were some specific search failures due to misspelling (one searcher had 0 Precision and Recall for one search due to this).

Analysis of the mean and standard deviation of precision and recall over all searches for each searcher and system showed a considerable range of performance within the searchers at Berkeley. In the following section we will examine the characteristics of the searchers as reported in the questionnaires administered during the experiments. Figure 4 summarizes the average aspectual precision and recall for each of the systems participating in the TREC-7 Interactive Track.

System	T							opic Number	Total
	352i	353i	357i	362i	365i	366i	387i	392i	
a	0.4418	0.3860	0.4870	0.5913	0.9168	0.9168	0.8125	0.6110	0.6454
b	0.7448	0.3483	0.4125	0.7500	0.9018	0.8875	0.7375	0.8310	0.7017
C	**0.7993**	**0.2145**	**0.6368**	**0.7058**	**0.9793**	**0.9375**	**0.8458**	**0.8068**	**0.7407**
clus	0.4355	0.3594	0.5514	0.6773	0.8750	0.4291	0.5938	0.6276	0.5686
irisa	0.7448	0.4333	0.5715	0.7500	0.7223	1.0000	0.9063	0.8088	0.7421
irisp	0.7333	0.1750	0.5533	0.6368	1.0000	0.9168	0.7500	0.7085	0.6842
iriss	0.6745	0.6021	0.5803	0.6194	0.8875	0.8611	0.7884	0.7303	0.7179
J24	0.6250	0.5178	0.5083	0.5568	0.9375	0.8750	0.6615	0.7315	0.6767
list	0.2811	0.1916	0.4200	0.5209	0.9584	0.3750	0.8084	0.7540	0.5387
MB	0.8831	0.4876	0.4164	0.7249	0.7524	0.9679	0.6494	0.6020	0.6855
MR	0.6342	0.4309	0.2857	0.7056	1.0000	1.0000	0.7944	0.7190	0.6962
ok_noRF	0.9018	0.4984	0.3033	0.4981	0.8840	0.8750	0.7710	0.4874	0.6524
ok_withRF	0.8578	0.5865	0.3818	0.3558	0.7520	1.0000	0.8335	0.7475	0.6893
RUINQ-G	0.5844	0.5109	0.3618	0.4628	0.8814	0.7918	0.8854	0.6228	0.6357
RUINQ-R	0.5558	0.3160	0.4391	0.6055	0.8674	0.7889	0.7689	0.6819	0.6286
Z	**0.9500**	**0.4405**	**0.5030**	**0.6043**	**1.0000**	**0.8333**	**0.9333**	**0.7093**	**0.7467**
ZP	0.5863	0.4045	0.4285	0.8928	0.9688	0.9500	0.8873	0.8493	0.7459
zp_noRF	0.8875	0.2323	0.4303	0.7640	0.8750	0.7918	0.9015	0.8250	0.7134
Grand Total	0.6733	0.4090	0.4367	0.6324	0.8928	0.8354	0.7776	0.6916	0.6689

Table 4: Aspectual Precision by System and Query

System							T	opic Number	Total
	352i	353i	357i	362i	365i	366i	387i	392i	
a	0.1250	0.2728	0.3658	0.3540	0.8023	0.5355	0.4723	0.3403	0.4085
b	0.2413	0.1593	0.3080	0.2918	0.8543	0.5000	0.3888	0.2433	0.3733
C	**0.2768**	**0.1820**	**0.3080**	**0.3123**	**0.8543**	**0.5358**	**0.4443**	**0.4373**	**0.4188**
clus	0.0893	0.1024	0.2310	0.1459	0.6874	0.2144	0.1110	0.1840	0.2207
irisa	0.2858	0.1593	0.2888	0.2708	0.5208	0.2860	0.1943	0.2430	0.2811
irisp	0.2413	0.1138	0.3655	0.1873	0.7918	0.5358	0.2220	0.3403	0.3497
iriss	0.1653	0.2161	0.2791	0.2605	0.7186	0.3571	0.3193	0.2985	0.3268
J24	0.1875	0.2275	0.3080	0.1875	0.8750	0.2503	0.4445	0.5280	0.3760
list	0.0534	0.0683	0.2791	0.1354	0.6770	0.0715	0.2499	0.2846	0.2274
MB	0.3368	0.1950	0.2090	0.3274	0.6161	0.4490	0.4522	0.2401	0.3532
MR	0.3061	0.1949	0.1155	0.2737	0.7260	0.3776	0.3171	0.4603	0.3464
ok_noRF	0.3349	0.2161	0.2406	0.3020	0.8489	0.4288	0.4305	0.2675	0.3837
ok_withRF	0.4105	0.2275	0.3655	0.3123	0.8230	0.3930	0.3610	0.2848	0.3972
RUINQ-G	0.2619	0.2340	0.2909	0.2499	0.8288	0.3930	0.4723	0.2918	0.3817
RUINQ-R	0.2365	0.2427	0.2214	0.3333	0.7449	0.3651	0.4073	0.2560	0.3489
Z	**0.2233**	**0.2955**	**0.2695**	**0.2918**	**0.8230**	**0.4285**	**0.6113**	**0.4723**	**0.4269**
ZP	0.2768	0.1593	0.3465	0.1668	0.8543	0.4643	0.5555	0.4308	0.4068
zp_noRF	0.3838	0.1138	0.3080	0.1878	0.8333	0.3930	0.5000	0.4723	0.3990
Grand Total	0.2478	0.1875	0.2573	0.2592	0.7503	0.3749	0.3750	0.3238	0.3472

Table 5: Aspectual Recall by Systems and Query

4.3.1 User Characteristics

The administration of the interactive track followed the track guidelines with a single group of 8 participants. While none of the participants had used either the experimental (Cheshire II) or control (ZPRISE) systems in searching tasks, many had seen demonstrations of the experimental system. The searchers who participated in the study were volunteers drawn from the School of Information Management and Systems at UC Berkeley (a call for participation was sent to all students and faculty at SIMS and the first 8 volunteers were scheduled for search sessions. A pre-search questionnaire asked each participant for basic demographic information and educational background, as well as their experience with various types of search systems.

All of the participants, except one undergraduate, held college degrees (One held a PhD, two others were PhD students with previous undergraduate and graduate degrees, and the remaining 4 were Masters students in the SIMS program). Two of the participants (P1 and P3) had considerable experience in online searching on other systems, the other two had very limited experience with online systems. The range of search experience with various systems varied from over 20 years to less than one year. By far, the most frequently used search systems were the Web search services and the next most frequent were online catalogs. Exploratory correlation analyses were performed on all of the variables from the presearch questionnaires, combined with the matching search and system questionnaires with the per-search and search precision and recall information from the NIST evaluators. Not surprisingly, there were very few significant correlations found in the analysis and many of those were trivial (years of search experience is positively correlated with age). Somewhat more interesting was that search experience with online systems like Dialog and BRS was also significantly correlated with search experience. It appears that most recent searchers will be gaining their experience from the WWW and possibly from online library catalogs, and will probably not have experience (or as much experience) with traditional Boolean systems such as Dialog.

4.3.2 Per Search Results

Following each search the participants were given a questionnaire asking about familiarity with the search topic, how easy it was to start and conduct the search and whether the user was satisfied with the results.

The responses from each participant are included in the WWW version of this section noted above. All searchers found the search easier to do with the ZPRISE system than with the Cheshire II system. Similarly, analysis of the average responses to the "Are you satisfied with the results" question showed that the ZPRISE system is given higher marks than the Cheshire system. Analysis was also conducted of the average responses to the question "Are you familiar with this topic?" Here the responses show that the searchers where generally less familiar with the topics searched on the Cheshire system versus those on the ZPRISE system. Correlation analysis showed, however, no significant correlation between familiarity with a topic and either the ease of searching or the satifaction with search results. Satisfaction was however fairly strongly correlated with how easy it was to do the search task (Pearson's R=0.646, prob=0.0001). Interestingly, there was no significant correlation between any of the post-search questions and either Aspectual Precision or Aspectual Recall, but the signs of these correlations indicated some interesting items for further research. For example, a slight negative correlation was indicated between Precision and Recall and the user's confidence that they had identified all of the different instances for a topic.

4.3.3 Post-System Questions

The searches were conducted in blocks of 4 questions on each system. Following the searcher's interaction with a system, a post-system questionnaire was administered. This post-system questionnaire asked each searcher questions about how easy the system was to learn, use, and understand, and permitted comments on the features of the system.

Overall, the searchers found both system very easy to learn. The Cheshire system was marked down again on the "easy to use" question. From the comments, this appeared to be related to some missing features (e.g. Boolean AND but no NOT), and several searchers mentioned the need to scroll back to the beginning of a record to select it as relevant. Others (who used ZPRISE first) mentioned a preference for having the full- text document in a separate window. With responses on a scale from 1 (difficult) to 5 (extremely easy), the average "ease of use" for Cheshire II was 3.38 and the average for ZPRISE was 4.25.

4.3.4 Exit Questionnaire

After the completion of all searches an exit questionnaire was administered to the searchers. This questionnaire asked how well the searchers understood the task, whether it was similar to other seach task, how they would rank the systems in relation to each other, and what they liked and disliked about each system.

The searchers claimed to have a very good understanding of the search task (mean was 4.25), and they found the task similar to other searching tasks (mean of 3.63). They also found the systems somewhat different (mean of 3.37). In ranking the systems, 5 out of 8 ranked ZPRISE as easier to learn to use, while 7 out of 8 chose ranked it as easier to use. Curiously the searchers were evenly split (four each) on which they liked the best. One search commented that she would prefer Cheshire "for serious research" but found ZPRISE better suited to the TREC search tasks.

5 Conclusions and Acknowledgments

In our TREC-7 experiments for the ad-hoc task and cross-language track, Berkeley utilized our probabilistic document retrieval methods for all retrieval. In the ad-hoc task we experimented with discovery of the "best" Boolean query and merged Boolean and probabilistic retrieval. This provided some spectacular successes and seemed to provided some overall improvement. In the interactive track, it was impossible to draw any firm general conclusions from our small sample of searchers and searches. But it is obvious that an interface that is well adapted to the specific search task will tend to be preferred by searchers even if the underlying system produces better overall performance in terms of Recall, and comparable performance in terms of precision. As observed at TREC-6, the overall performance of the Cheshire II system was quite good, although it was not dramatically better than the control system on average. These results, as has often been noted in previous TREC interactive evaluations, tend to be highly influenced by individual behavior and search techniques (this is apparent in the differences between the searchers on the same questions and in the same systems). What seems apparent from the results of the questionnaires coupled with the Precision and Recall measures is that a generic interface can perform quite acceptably in the TREC tasks, even if it isn't particularly liked by the users, compared to another system that is better suited to the TREC tasks, but may not be as useful in other situations.

This research was supported by the Information and Data Management Program of the National Science Foundation under grant IRI-9630765 from the Database and Expert Systems program of the Computer and Information Science and Engineering Directorate. The original development of the Cheshire II system was sponsored by a College Library Technology and Cooperation Grants Program, HEA-IIA, Research and Demonstration Grant #R197D30040 from the U.S. Department of Education. Further development work on the Cheshire II project and system was supported as part of Berkeley's NSF/NASA/ARPA Digital Library Initiative Grant #IRI-9411334.

Both projects are currently being supported in part by DARPA (Department of Defense Advanced Research Projects Agency) under research contract N66001-97-C-8541, AO-F477.

References

[1] EasyTranslator Version 2.0. Transparent Language Inc.

[2] Globalink Power Translator Version 6.02. Globalink Inc.

[3] W. S. Cooper, A. Chen, and F. C. Gey. Full Text Retrieval based on Probabilistic Equations with Coeffici ents fitted by Logistic Regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.

[4] William S. Cooper, Aitao Chen, and Fredric C. Gey. Experiments in the probabilistic retrieval of full text documents. In *Text Retrieval Conference (TREC-3) Draft Conference Papers*, Gaithersburg, MD, 1994. National Institute of Standards and Technology.

[5] William S. Cooper, Fredric C. Gey, and Daniel P. Dabney. Probabilistic retrieval based on staged logistic regression. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24*, pages 198–210, New York, 1992. ACM.

[6] C.R. Palmer G. V. Cormack, C.L.A. Clarke and S.S.L. To. Passage-Based Refinement (MultiText Experiments for TREC-6). In D. K. Harman and Ellen Voorhees, editors, *The Sixth Text REtrieval Conference (TREC-6), NIST Special Publication 500-240*, pages 303–319, August 1998.

[7] F. C. Gey and A. Chen. Phrase Discovery for English and Cross-language Retrieval at TREC-6. In D. K. Harman and Ellen Voorhees, editors, *The Sixth Text REtrieval Conference (TREC-6), NIST Special Publication 500-240*, pages 637–647, August 1998.

[8] Marti Hearst. Improving Full-Text Precision on Short Queries using Simple Constraints. In *The Fifth Annual Symposium on Document Analysis and Information Retrieval SDAIR), Las Vegas, Nevada*, April 1996.

[9] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression.* John Wiley & Sons, New York, 1989.

[10] Ray R. Larson and Jerome McDonough. Cheshire II at TREC 6: Interactive probabilistic retrieval. In Donna Harman and Ellen Voorhees, editors, *TREC 6 Proceedings (Notebook)*, pages 405–415, Gaithersburg, MD, 1997. National Institute of Standards and Technology.

[11] Ray R. Larson, Jerome McDonough, Paul O'Leary, Lucy Kuntz, and Ralph Moon. Cheshire II: Designing a next-generation online catalog. *Journal of the American Society for Information Science*, 47(7):555–567, July 1996.

[12] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.

[13] SYSTRAN:. *http://babelfish.altavista.digital.com/.*