Comparing User-assisted and Automatic Query Translation

Daqing He[1], Jianqiang Wang[2], Douglas W. Oard[2], Michael Nossal[1]

[1] Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
{daqingd,nossal}@umiacs.umd.edu
[2] College of Information Studies & Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
{wangjq,oard}@glue.umd.edu

Abstract. For the 2002 Cross-Language Evaluation Forum Interactive Track, the University of Maryland team focused on query formulation and reformulation. Twelve people performed a total of forty eight searches in the German document collection using English queries. Half of the searches were with user-assisted query translation, and half with fully automatic query translation. For the user-assisted query translation condition, participants were provided two types of cues about the meaning of each translation: a list of other terms with the same translation (potential synonyms), and a sentence in which the word was used in a translation-appropriate context. Four searchers performed the official iCLEF task, the other eight searched a smaller collection. Searchers performing the official task were able to make more accurate relevance judgments with user-assisted query translation for three of the four topics. We observed that the number of query iterations seems to vary systematically with topic, system, and collection, and we are analyzing query content and ranked retrieval measures to obtain further insight into these variations in search behavior.

1 Introduction

Interactive Cross Language Information Retrieval (CLIR) is an iterative process in which searcher and system collaborate to find documents that satisfy an information need, regardless of whether they are written in the same language as the query. Humans and machines bring complementary strengths to this process. Machines are excellent at repetitive tasks that are well specified; humans bring creativity and exceptional pattern recognition capabilities. Properly coupling these capabilities can result in a synergy that greatly exceeds the ability of either human or machine alone. The design of the fully automated components to support cross-language searching (e.g., structured query translation and ranked retrieval) has been well researched, but achieving true synergy requires that the machine also provide tools that will allow its human partners to exercise their skills to the greatest possible degree. Such tools are the focus of our work in

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005		**00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Comparing User-assisted and Automatic Query Translation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defense Advanced Research Projects Agency,3701 North Fairfax Drive,Arlington,VA,22203-1714**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**15**	

the Cross-Language Evaluation Forum's (CLEF) interactive track (iCLEF). In 2001, we began by exploring support for document selection [5]. This year, our focus is on query formulation.

Cross-language retrieval techniques can generally be classified as query translation, document translation, or interlingual designs [2]. We adopted a query translation design because the query translation stage provides an additional interaction opportunity not present in document translation based systems. Our searchers first formulate a query in English, then the system translates that query into the document language (German, in our case). The translated query is used to search the document collection, and a ranked list of document surrogates (first 40 words, in our case) is displayed. The searcher can examine individual documents, and can optionally repeat the process by reformulating the query. Although there are only three possible interaction points (query formulation, query translation, and document selection), the iterative nature of the process introduces significant complexity. We therefore performed extensive exploratory data analysis to understand how searchers employ the systems that we provided.

Our study was motivated by the following questions:

1. What strategies do searchers apply when formulating their initial query and when reformulating that query? In what ways do their strategies differ from those used in monolingual applications? How do individual differences in subject knowledge, language skills, search experience, and other factors affect this process?
2. What information do searchers need when reformulating their query, and how do they obtain that information?
3. Can searchers find documents more effectively if we give them some degree of control over the query translation process? Do searchers prefer to exercise control over the query translation process? What reasons do they give for their preference?
4. What measures can best illuminate the effect of interactive query reformulation on retrieval effectiveness?

These questions are, of course, far to broad to be answered completely by any single experiment. For the experiments reported in this paper, we chose to provide our searchers with two variants on a single retrieval system, one with support for interaction during query translation (which we call "manual"), and the other with fully automatic query translation (which we call "auto"). This design allowed us to test a hypothesis derived from our third question above. We relied on observations, questionnaires, semi-structured interviews, and exploratory data analysis to augment the insight gained through hypothesis testing, and to begin our exploration of the other questions.

In the next section, we describe the design of our system. Section 3 then describes our experiment, and Section 4 presents the results that we obtained. Section 5 concludes the paper with a brief discussion of future work.

2 System Design

In this section, we describe the resources that we used, the design of our cross-language retrieval system, and our user interface design.

2.1 Resources

We chose English as the query language and German as the document language because our population of potential searchers was generally skilled in English but not German. The full German document collection contained 71,677 news stories from the Swiss News Agency (SDA) and 13,979 new stories from Der-Spiegel. We used the German-to-English translations provided by the iCLEF organizers for construction of document surrogates (for display in a ranked list) and for display of full document translations (when selected for viewing by the searcher). The translations were created using Systran Professional 3.0.

We obtained a German-English bilingual term list from the Chemnitz University of Technology [3], and used the German stemmer from the "snowball" project [4]. Our Keyword in Context (KWIC) technique requires parallel (i.e., translation-equivalent) German/English texts – we obtained those from the Foreign Broadcast Information Service (FBIS) TIDES data disk, release 2.

2.2 CLIR System

We used the InQuery text retrieval system (version 3.1p1) from the University of Massachusetts, along with locally implemented extensions to support cross-language retrieval between German and English. We used Pirkola's structured query technique for query translation [4], which aggregates German term frequencies and document frequencies separately before computing the weight for each English query term. This tends to suppress the contribution to the ranking computations of those English terms that have at least one translation that is a common German word (i.e., that occurs in many documents). For the automatic condition, all known translations were used. For the manual condition, only translations selected by the searcher were used. We employed a backoff translation strategy to maximize the coverage of the bilingual term list [3]. If no translation was found for the surface form of an English term, we stemmed the term (using the Porter stemmer) and tried again. If that failed, we also stemmed the English side of the bilingual term list and tried a third time. If that still failed, we treated the untranslated term as its own translation in the hope that it might be a proper name.

2.3 User Interface Design

For our automatic condition, we adopted an interface design similar to that of present Web search engines. Searchers entered English query terms in an one-line text field, based on their understanding of a full CLEF topic description

[3] http://dict.tu-chemnitz.de/
[4] http://snowball.sourceforge.net

(title, description, and narrative). We provided that topic description on paper in order to encourage a more natural query formulation process than might have not been the case if cut-and-paste from the topic description were available. When the search button was clicked, a ranked list of document surrogates was displayed below the query field, thus allowing the query to serve as context when interpreting the ranked list. Ten surrogates were displayed simultaneously as a page, and up to 10 pages (in total 100 surrogates) could be viewed by clicking "next" button. Our surrogates consisted of the first 40 words in the TEXT field of the translated document. English words in the surrogate that shared a common stem with any query term (using the Porter stemmer) were highlighted in red. See Figure 1 for an illustration of the automatic user interface.

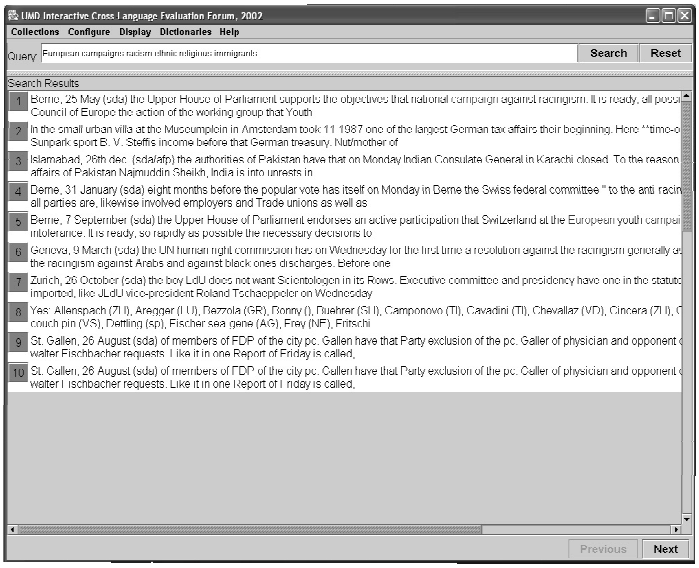

Fig. 1. User interface, automatic condition.

Each surrogate is labeled with a numeric rank (1, 2, 3, . . .), which is displayed as a numbered button to the left of the surrogate. If the searcher selected the button, the full text of that document would be displayed in a separate window, with query terms highlighted in the same manner. In order to provide context, we repeated the numeric rank and the surrogate at the top of the document examination window. Figure 2 illustrates a document examination window.

We collected three types of information about relevance judgments. First, searchers could indicate whether the document was not relevant ("N"), somewhat relevant ("S"), or highly relevant ("H"). A fourth value, "?" (indicating unjudged), was initially selected by the system. Second, searchers could indicate their degree of confidence in their judgment as low ("L"), medium ("M"), or high ("H"), with a fourth value ("?") being initially selected by the system.

Both relevance judgments and confidence values were recorded incrementally in a log file. Searchers could record relevance judgments and confidence values in either the main search window or in a document examination window (when that window was displayed). Finally, we recorded the times at which documents were selected for examination and the times at which relevance judgments for those documents were recorded. This allowed us to later compute the (approximate) examination time for each document. For documents that were judged without examination (e.g., based solely on the surrogate), we assigned zero as the examination time.

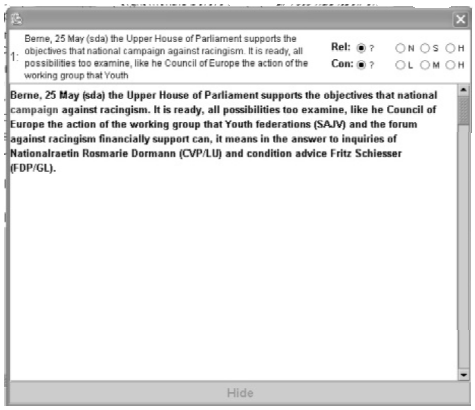

Fig. 2. Document examination window.

For the manual interface, we used a variant of the same interface with two additional items: 1) term-by-term control over the query translation process, and 2) a summary of the translations chosen for all query terms. We used a tabbed pane to allow the user to examine alternative translations for one English query term at a time. Each possible translation was shown on a separate line, and a check-box to the left of each line allowed the user to deselect or reselect that translation. All translations were initially selected, so the manual and automatic conditions would be identical if the user did not deselect any translation.

Since we designed our interface to support searchers with no knowledge of German, we provided cues in English about the meaning of each German translation. For these experiments, searchers were able to view two types of cues: (1) back translation, and (2) Keyword In Context (KWIC). Each was created automatically, using techniques described below. Searchers were able to alternate between the two types of cues using tabs. The query translation summary area provided additional context for interpretation of the ranked list, simultaneously showing all selected translations (with one back translation each). In order to emphasize that two steps were involved (query translation, followed by search), we provided both “translate query” and “search” buttons. All other functions

were identical to the automatic condition. Figure 3 illustrates the user interface for the manual condition.

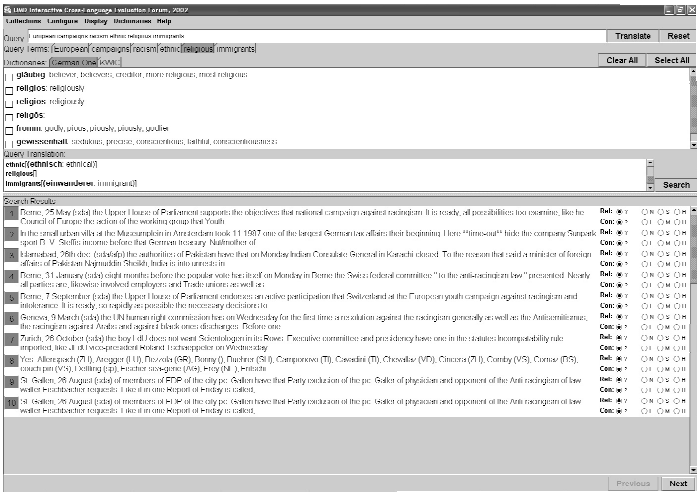

Fig. 3. User interface, manual condition.

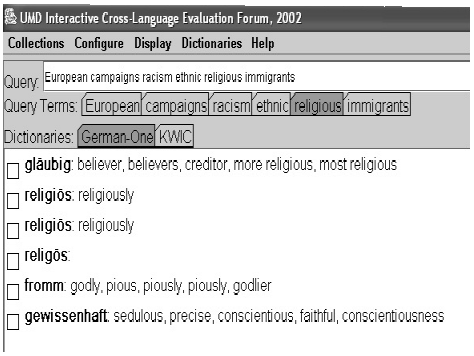

Fig. 4. Back translations of "religious."

Back Translation Ideally, we would prefer to provide the searcher with English definitions for each German translation alternative. Dictionaries with these types of definitions do exist for some language pairs (although rights management considerations may limit their availability in electronic form), but bilingual term lists are much more easily available. What we call "back translations" are English

terms that share a specific German translation, something that we can determine with a simple bilingual term list. For example, the English word `religious` has several German translations in the term list that we used, two of which are `fromm` and `gewissenhaft`. Looking in the same term list for cues to the meaning of `fromm`, we see that it can be translated into English as `religious, godly, pious, piously,` or `godiler`. Thus `fromm` seems to clearly correspond to the literal use of `religious`. By contrast, `gewissenhaft`'s back translations are `religious, sedulous, precise, conscientious, faithful,` or `conscientiousness`. This seems as if it might correspond with a more figurative use of `religious`, as in "he rode his bike to work religiously." Of course, many German translations will themselves have multiple senses, so detecting a reliable signal in the noisy cues provided by back translation sometimes requires common-sense reasoning. Fortunately, that is a task for which human are uniquely well suited. The original English term will always be its own back translation, so we supress its display. Sometimes this results in an empty (and therefore uninformative) set of back translations. Figure 4 shows the back translation display for "religious" in our manual condition.

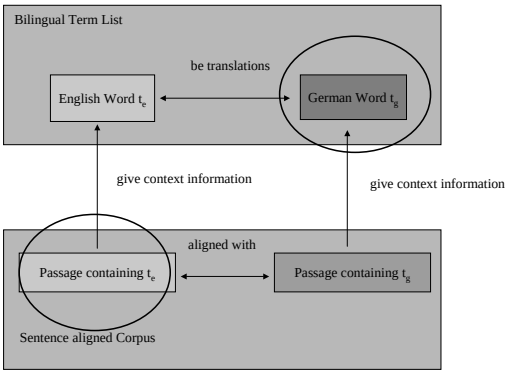

Fig. 5. Constructing cross-language KWIC using a sentence-aligned parallel corpus.

Keyword in Context One way to compensate for the weaknesses of back translation is to draw additional evidence from examples of usage. In keeping with the common usage in monolingual contexts [1], we call this approach "keyword in context" or "KWIC." For each German translation of an English term, our goal is to find a brief passage in which the English term is used in a manner appropriate to the translation in question. To do this, we started with a collection of document pairs that are translations of each other. We used German news stories that had previously been manually translated into English by the Foreign Broadcast Information Service (FBIS) and distributed as a standard research corpus. We segmented the FBIS documents into sentences using

rule-based software based on punctuation and capitalization patterns, and then produced aligned sentence pairs using the GSA algorithm (which uses dynamic programming to discover a plausible mapping of sentences within a paired documents based upon known translation relationships from the bilingual term list, sentence lengths and relative positions in each documents). We presented the entire English sentence, favoring the shortest one if multiple sentence pairs contained the same English term.[5]

Formally, let t_e be an English term for which we seek an example of usage, and let t_g be the German translation from the bilingual term list that is of interest. Let S_e and S_g be the shortest pair of sentences that contain t_e and t_g respectively. We then present S_e as the example of usage for translation t_g. Figure 5 illustrates this process.

3 Experiment Design

Our experiment is designed to test the utility of user-assisted query translation in an interactive cross-language retrieval system. We were motivated to explore this question by two potential benefits that we foresaw:

- The effectiveness of ranked retrieval might be improved if a more refined set of translations for key query terms were known.
- The searcher's ability to employ the retrieval system might be improved by providing greater transparency for the query translation process.

Formally, we sought to reject the null hypotheses that there is no difference between the $F_{\alpha=0.8}$ achieved using the automatic and manual systems. The F measure is an outcome measure, however, and we were also interested in understanding process issues. We used exploratory data analysis to improve our understanding how the searchers used the cues we provided.

3.1 Procedure

We followed the standard protocol for iCLEF 2002 experiments. Searchers were sequentially given four topics (stated in English), two for use with the manual system and two for use with the automatic system. Presentation order for topics and system was varied systematically across searchers as specified in the track guidelines. After an initial training session, they were given 20 minutes for each search to identify relevant documents using the radio buttons provided for that purpose in our user interface. The searchers were asked to emphasize precision over recall (by telling them that it was more important that the document that they selected be truly relevant than that they find every possible relevant document). We asked each searcher to fill out brief questionnaires before the first

[5] We did not highlight the query term in current version due to time constraints. Another limitation of current implementation is that a briefer passage may serve our purpose better in some cases.

search (for demographic data), after each search, and after using each system. Each searcher used the same system at a different time, so we were able to observe each individually and make extensive observational notes. We also conducted a semi-structured interview (in which we tailored our questions based on our observations) after all searches were completed.

We conducted a pilot study with a single searcher (umd01) to exercise our new system and refine our data collection procedures. Eight searchers (umd02-umd09) then performed the experiment using the eight-subject design specified in the track guidelines [6]. While preparing our results for submission, we noticed that no SDA document appeared in any ranked list. Investigation revealed that InQuery had failed to index those documents because we had not configured the SGML parsing correctly for that collection. We therefore corrected that problem, recruited four new searchers (umd10-umd13), and repeated the experiment, this time using the four-subject design specified in the track guidelines.

We submitted all twelve runs for use in forming relevance pools, but designated the second experiment as our official submission because the first experiment did not comply with one requirement of the track guidelines (the collections to be searched). Our results from the first experiment are, however, interesting for several reasons. First, it turned out that topic 3 had no relevant documents in the collection searched in the first experiment.[7] This happens in real applications, of course, but the situation is rarely studied in information retrieval experiments because the typical evaluation measures are unable to discriminate between systems when no relevant documents are exist. Second, the number of relevant documents for the remaining three topics was smaller in the first experiment than the second. This provided an opportunity to study the effect of collection characteristics on searcher behavior.

For convenience, we refer to the first experiment as the *small collection* experiment, and the second as the *large collection* experiment.

3.2 Measures

We computed the following measures in order to gain insight into search behavior and search results:

- $F_{\alpha=0.8}$, as defined in the track guidelines (with "somewat relevant" documents treated as not relevant). We refer to this condition as "strict" relevance judgments. This value was computed at the end of each search session.
- $F_{\alpha=0.8}$, but with "somewat relevant" documents treated as relevant. We refer to this condition as "loose" relevance judgments. This vale was also computed for each session.
- Mean uninterpolated Average Precision (MAP) for the ranked list returned by each iteration of a search process.

[6] http://terral.lsi.uned.es/iclef/2002/

[7] In this paper, we number the topics 1, 2, 3, and 4 in keeping with the track guidelines. These correspond to CLEF topic numbers c053, c065,c056 and c080, respectively.

– A variant of MAP in which documents already marked as "highly relevant" are placed at the top of the ranked list (in an arbitrary order). We refer to this measure as "MAP-S" (for "strict").
– A second variant of MAP in which documents already marked as "highly relevant" or "somewhat relevant" are placed at the top of the ranked list (in an arbitrary order). We refer to this measure as "MAP-L" (for "loose").
– A third variant of MAP in which only the documents statisfying the two conditions – 1) they are already marked as "highly relevant" by the subject; 2) they are the real relevant documents according to "ground truth" – are placed at the top of the ranked list (in an arbitrary order). We refer to this measure as "MAP-R" (for "real").
– The total examination time (in seconds) for each document, summed over all instances of examination for the same document. If the full text of a document was never examined, an examination time of zero was recorded.
– The total number of query iterations for each search.

The set oriented measures (strict and loose F) are deigned to characterize end-to-end task performance using the system. The rank-oriented measures (MAP, MAP-S, MAP-L and MAP-R) are designed to offer indirect insight into the query formulation process by characterizing the effect of a query based on the density of relevant documents near the top of the ranked list produced for that query (or for queries up through that iteration by either viewing from the point of the subject's own sense of performance, in the case of MAP-S and MAP-L, or viewing from the actual performance, in the case of MAP-R). Examination time is intended for use in conjunction with relevance judgment categories, in order to gain some insight into the relevance judgment process. We have not yet finished our trajectory analysis or the analysis of examination duration, so in this paper we report results only for the final values of $F_{\alpha=0.8}$ and for the number of iterations.

4 Results

4.1 Searchers

Our searcher population was relatively homogeneous. Specifically, they were:

Affiliated with a university. Every one of our searchers was a student, staff member or faculty member at the University of Maryland.

Highly educated. Ten of the 12 searchers are either enrolled in a Masters degree program or had earned a Masters degree or higher. The remaining two were undergraduate students, and they are both in the small collection experiment.

Mature. The average age over all 12 searchers was 31, with the youngest being 19 and the oldest being 43. The average age of the four searchers in the large collection experiment was 32.

Mostly female. There were three times as many female searchers as males, both overall and in the large collection experiment.

Experienced searchers. Six of the 12 searchers held degrees in library science. The searchers reported an average of about 6 years of on-line searching experience, with a minimum of 4 years and maximum of 10 years. Most searchers reported extensive experience with Web search services, and all reported at least some experience searching computerized library catalogs (ranging from "some" to "a great deal"). Eleven of the 12 reported that they search at least once or twice a day. The search experience data for the four participants in the large collection experiment was slightly greater than for the 12 searchers as a whole.

Not previous study participants. None of the 12 subjects had previously participated in a TREC or or iCLEF study.

Inexperienced with machine translation. Nine of the 12 participants reported never having used any machine translation software or free Web translation service. The other 3 reported "very little experience" with machine translation software or services. The four participants in the large collection experiment reported the same ratio.

Native English speakers. All 12 searchers were native speakers of English.

Not skilled in German. Eight of the 12 searchers reported no reading skills in German at all. Another 3 reported poor reading skills in German, and one (umd12) reported good reading skill in German. Among the four searchers in the large collection experiment, 3 reported no German skills, with the fourth reporting good reading skills in German.

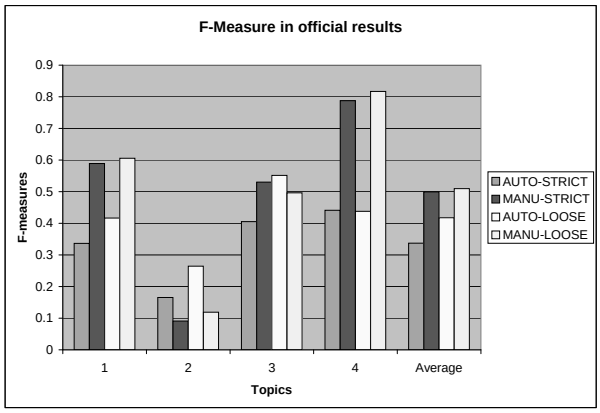

Fig. 6. $F_\alpha = 0.8$, large collection, by condition and topic.

4.2 Large Collection Experiment

Our official results on the large collection experiment found that the manual system achieved a 48% larger value for $F_{\alpha=0.8}$ than the automatic system (0.4995

vs. 0.3371). However, the difference is not statistically significant, and the most likely reason is the samll sample size. The presence of a searcher with good reading skills in German is also potentially troublesome given the hypothesis that we wished to test. We have not yet conducted searcher-by-searcher analysis to determine whether searcher umd12 exhibited search behaviors markedly different from the other 11 searchers. For contrast, we recomputed the same results with loose relevance. In that case, the searchers in our large collection experiment achieved a 22% increase in $F_{\alpha=0.8}$ over the automatic system (0.5095 vs. 0.4176).

As Figure 6 shows, the manual system achieved the largest improvements for topics 1 (Genes and Diseases) and 4 (Hunger Strikes) with strict relevance, but the automatic system actually outperformed the manual system on topic 2 (Treasure Hunting). Loose relevance judgments exhibited a similar pattern. Searchers that were presented with topic 2 in the manual condition reported (in questionnaire) that it was more difficult to identify appropriate translations for topic 2 than for any other topic, and searchers generally indicated that they were less familiar with topic 2 than with other topics. We have not yet completed our analysis of observational notes, so we are not able to say whether this resulted in any differences in search behavior. But it seems likely that without useful cues, searchers removed translations that would have been better off keeping. If confirmed through further analysis, this may have implications for user training.

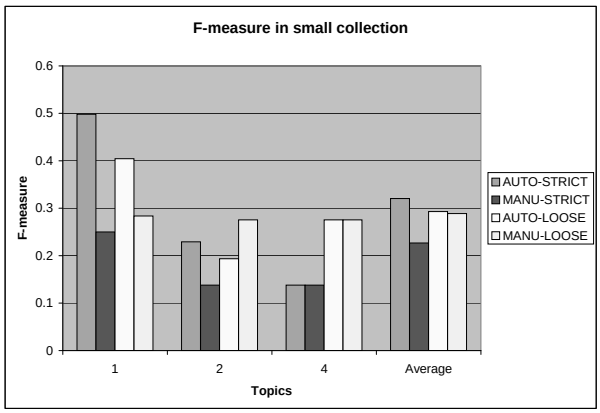

Fig. 7. $F_\alpha = 0.8$, small collection group, by condition.

4.3 Small Collection Experiment

The results of the small collection experiment shown in Figure 7 are quite different. The situation is reversed for topic 1, with automatic now outperforming manual, and topic 4 no longer discriminates between the two systems.[8] Overall,

[8] Topic 3, with no relevant documents in the small collection, is not shown.

the manual and automatic systems could not be distinguished using loose relevance (0.2889 vs 0.2931), but the automatic system seemed to do better with strict relevance (0.2268 vs 0.3206). Again, we did not find that the difference is statistically significant. The data that we have analyzed does, however, seem to suggest that our manual system is better suited to cases in which there are a substantial number of relevant documents. We plan to use this question to guide some of our further data analysis.

4.4 Subjective Assessment

We analyzed questionnaire data and interview responses in an effort to understand how participants employed the systems and to better understand their impressions about the systems. Questionnaire responses are on a 1-5 scale (with 1 being "not at all," and 5 being "strongly agree").

Searchers in the large collection experiment reported that the manual and automatic systems were equally easy to search with (average 3.5), but searchers in the small collection experiment reported that the automatic system was easier to use than the manual system (3.4 vs. 2.75).

Searchers in the large collection experiment reported an equal need to reformulate their initial queries with both systems (average 3.25), but searchers in the small collection experiment reported that this was somewhat less necessary with the automatic system (3.9 vs. 4.1). One searcher, umd07 reported that it was "extremely necessary" to reformulate queries with both systems. We notice from his/her answers to our open questions that he/she thought the query translations were "usually very poor," and he/she would like both systems support Boolean queries, proximity operators and truncations so that "noise" could be removed.

Searchers in the large collection experiment reported that they were able to find relevant documents more easily using the manual system than the automatic system (4.0 vs. 3.5), but searchers in the small collection experiment had the opposite opinion (2.6 vs. 3.0).

For questions unique to the manual system, the large collection group reported positive reactions to the usefulness of user-assisted query translation (with everyone choosing a value of 4). They generally felt that it was possible to identify unintended translations (an average of 3.5), and that and most of the time the system provided appropriate translations (average of 3.9).

Most participants reported that they were not familiar with the topics, with topic 3 (European Campaigns against Racism) having the most familiarity, and topics 1 and 2 having the least.

4.5 Query iteration analysis

We determined the number of iterations for each search through log file analysis. In the large collection experiment, searchers averaged 9 query iterations per search across all conditions. Topic 2 had the largest number of iterations (averaging 16), topic 4 had the fewest (averaging 6). Topics 1 and 2 exhibited little

difference in the average number of iterations across systems, but topics 3 and 4 had substantially fewer iterations with the manual system. In the small collection experiment, searchers performed substantially more iterations per search than that in the large collection experiment, averaging 13 iterations per search across all conditions. Topic 2 again has the greatest number of iterations (averaging 16), while topic 1 had the fewest (averaging 8).

4.6 The effect of the number of relevant documents

The unexpected problem with indexing the SDA collection reduced the number of searchers that contributed to our official results, but it provided us with an extra dimension for our analysis. Searchers in the large collection and small collection experiments were generally drawn from the same population, were given the same topics, used the same systems, and performed the same tasks. The main difference is the nature of the collection that they searched, and in particular the number of relevant documents that were available to be found. Summarizing the results above from this perspective, we observed the following differences between the two experiments:

- Objectively, searchers seemed to achieve a better outcome measure with the manual system in the large collection experiment, but they seemed to do better with the automatic system in the small collection experiment.
- Subjectively, searchers preferred using the manual system in the large collection experiment, but they preferred the automatic system in the small collection experiment.
- Examining search behavior, we found that the average number of query refinement iterations per search was inversely correlated with the number of relevant documents.

We have not yet finished our analysis, but the preponderance of the evidence that is presently available suggests that collection characteristics may be an important variable in the design of interactive CLIR systems. We believe that this factor should receive attention in future work on this subject.

5 Conclusion and future work

We focused on supporting user participation in the query translation process, and tested the effectiveness of two types of cues—*back translation* and *keyword in context* in an interactive CLIR application. Our preliminary analysis suggests that together these cues can sometimes be helpful, but that the degree of utility that is obtained is dependent on the characteristics of the topic, the collection, and the available translation resources.

Our experiments suggest a number of promising directions for future work. First, mean average precision is a commonly reported measure for the quality of a ranked list (and, by extension, for the quality of the query that led to the creation of that ranked list). We have found that it is difficult to draw

insights from MAP trajectories (variations across sequential query refinement iterations), in part because we do not yet have a good way to describe the strategies that a searcher might employ. We are presently working to characterize these strategies in a useful way, and to develop variants of the MAP measure (three of which were described above) that may offer additional insight. Second, our initial experiments with using KWIC for user-assisted query translation seem promising, but there are several things that we might improve. For example, it would be better if we could find the examples of usage in a comparable corpus (or even the Web) rather than a parallel corpus because parallel corpora are difficult to obtain. Finally, we observed far more query reformulation activity in this study than we had expected to see. Our present system provides some support for reformulation by allowing the user to see which query term translations are being used in the search. But we do not yet provide the searcher with any insight into the second half of that process—which German words correspond to potentially useful English terms that are learned by examining the translations? If we used the same resources for document translation as for query translation, this might not be a serious problem. But we don't, so it is an issue that we need to think about how to support.

The CLEF interactive track had proven to be an excellent source of insight into both system design and experiment design. We look forward to next year's experiments!

Acknowledgments

The authors would like to thank Julio Gonzalo and Fernando López-Ostenero for their tireless efforts to coordinate iCLEF. This work has been supported in part by DARPA cooperative agreement N660010028910.

References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
2. Douglas W. Oard and Anne Diekema. Cross-Language Information Retrieval. *Annual Review of Information Science and Technology*, 33:223–256, 1998.
3. Douglas W. Oard, Gina-Anne Levow, and Clara I. Cabezas. CLEF Experiments at Maryland: Statistical Stemming and backoff translation. In C. Peters, editor, *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000*, pages 176–187, Lisbon, Portugal, 2000.
4. Ari Pirkola. The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998.
5. Jianqiang Wang and Douglas W. Oard. iCLEF 2001 at Maryland: Comparing Word-for-Word Gloss and MT. In C. Peters, M. Braschler, J. Gonzalo, and Kluck M, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, pages 336–354, Darmstadt, Germany, 2001.