iCLEF 2001 at Maryland: Comparing Term-for-Term Gloss and MT

Jianqiang Wang and Douglas W. Oard

Human Computer Interaction Laboratory
College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA
{oard,wangjq}@glue.umd.edu,
WWW home page: http://www.glue.umd.edu/~oard/

Abstract. For the first interactive Cross-Language Evaluation Forum, the Maryland team focused on comparison of term-for-term gloss translation with full machine translation for the document selection task. The results show that (1) searchers are able to make relevance judgments with translations from either approach, and (2) the machine translation system achieved better effectiveness than the gloss translation strategy that we tried, although the difference is not statistically significant. It was noted that the "somewhat relevant" category was used differently by searchers presented with gloss translations than with machine translations, and some reasons for that difference are suggested. Finally, the results suggest that the F measure used in this evaluation is better suited for use with topics that have many known relevant documents than those with few.

1 Introduction

In the process of interactive cross-language information retrieval (CLIR), there are two points where interaction with the searcher is possible: query formulation and document selection. The focus of this paper is on the interactive document selection task. Ranked retrieval systems nominate promising documents for examination by the user by placing them higher in a ranked list. The searcher's task is then to examine those documents and select the ones that help to meet their information need. The query formulation process and the actual use of the documents selected by the user is outside the scope of the work reported in this paper. Focusing on one aspect of the problem in this way makes it possible to gain insight through the use of metrics that are appropriate for document selection, a well-studied problem in other contexts.

One important use for CLIR systems is to help searchers find information that is written in a language with which they are not familiar. In such an application, the query would be posed in a language for which the user has an adequate active (i.e., writing) vocabulary, and the document selection process would be performed in a language for which the searcher has at least an adequate passive

(i.e., reading) vocabulary. Since we have assumed that the document(s) being sought are not expressed in such a language, some form of translation is required.

We view translation as a user interface design challenge, in which the goal is to provide the user with the information needed to perform some task—in this case document selection. There has been an extensive effort to develop so-called "Machine Translation" (MT) systems to produce (hopefully) fluent and accurate translation of every language that is presently studied at the Cross-Language Evaluation Forum (CLEF) into English. No such systems yet exist for most of the world's languages, however, and the cost of building a sophisticated MT system for every written language would indeed be staggering. This is an important challenge, since a substantial portion of the world's knowledge is presently recorded in English, and the vast majority of the world's people cannot even find that information. Supporting search by users that know only a lesser-developed language is only one of many capabilities that will be needed if we are to address what has been called the "digital divide" on a global scale. But it is one that we believe could be addressed with emerging broad-coverage language technologies. We therefore have chosen to use this first interactive CLEF (iCLEF) evaluation to begin to explore that question.

We have identified three factors that affect the utility of translation technology for the document selection task: accuracy, fluency, and focus. By "accuracy" we mean the degree to which a translation reflects the intent of the original author. Both lexical selection (word choice) and presentation order can affect accuracy.[1] By "fluency" we mean the degree to which a translation can be used quickly to achieve the intended purpose (in this case, document selection). Again, both lexical selection and presentation order can affect fluency.[2] By focus, we mean the degree to which the reader's attention can be focused on the portions of a translated document that best support the intended task–in this case the recognition of relevant documents from among those nominated by the system. Presentation of summaries and highlighting query terms in the retrieved documents are typical examples of focus. Our intuition suggests that accuracy is essential for the document selection task, but that there is a tradeoff between fluency and focus, with lower fluency being acceptable if effective focus mechanisms are provided. The iCLEF evaluation was well timed to allow us to begin to explore these questions.

For iCLEF, we chose to compare MT with a one-best term-by-term gloss translation technique that we had originally developed to demonstrate the degree of translation quality that could be achieved for resource-poor languages. We had already adapted this system to support controlled user studies for some exploratory work on interactive document selection in the CLIR track of the 2000 Text Retrieval Conference (TREC) [4], so only minor modifications were needed to conform to the iCLEF requirements. We obtained a number of interesting results, including:

[1] Consider the case of "Harry hit Tom" and "Tom hit Harry" to see why presentation order can be an accuracy issue.

[2] For example, "Tom hitting by Harry" is understandable, but disfluent.

- Searchers are able to make some useful relevance judgments with either type of translation
- MT achieved better effectiveness than gloss translation, although the difference was not statistically significant
- The "somewhat relevant" category was used differently by participants in our experiment depending on whether MT or gloss translations were being examined.
- The F_α effectiveness measure does not seem to be well suited for use with topics that have few relevant documents.

The remainder of the paper is organized as follows. Section 2 provides an overview of the iCLEF experiment design. Section 3 then describes the design and implementation of our system, details of the experiment procedure, and a description of the characteristics of the participants in our experiment. Section 4 presents the results, drawing on both quantitative and qualitative methods, and raises some experiment design issues. Finally, Section 5 concludes the paper.

2 Background

Over the past decade, research on CLIR has focused on development and evaluation of automatic approaches for ranking documents in a language different from that of the query. Present fully automatic techniques can do this fairly well, performing at perhaps 80% of what can be achieved by a monolingual information retrieval system under similar conditions when measured using mean average precision [3]. Ranking documents is only one step in a search process, however; some means of selecting documents from that list is needed. One possible strategy would be to build an automatic classifier that could make a sharp decision about whether each document is relevant or not. Such an approach would have problems, however, since users often don't express their information needs clearly. Indeed, they may not even *know* their information needs clearly at the outset of a search session. For this reason, ranked retrieval systems are often used interactively, with the user browsing the ranked list and selecting interesting documents. Research on interactive retrieval strongly suggests that people are quite good at this task, performing quite well even when using ranked lists produced by systems that are well below the current state-of-the-art [1]. It is an open question, however, whether a similar strategy would be effective if automatically produced translations of otherwise unreadable documents would be sufficient to obtain a similar effect in interactive CLIR applications. The goal of the iCLEF evaluation is to bring together a research community to explore that question [2].

The principal objective of the first iCLEF evaluation was to develop an experiment design that could yield insight into the effectiveness of alternative techniques for supporting cross-language document selection. Participating sites could choose from two tasks: Selection of French documents or selection of English documents. We chose to work on selection of French documents since knowledge of French among the pool of possible participants in our experiment was

more limited than knowledge of English. The French test collection contained four search topics for use in the experiment, plus a fifth practice topic. For each topic, the following resources were provided:

- An English topic description, consisting of title, description, and narrative that served as a basis for the CLIR system's query,
- A ranked list of the top 50 documents produced automatically by a CLIR system using an English query,
- The original French version of each document, and
- An English translation of each document that was produced using Systran Professional 3.0.

The four topics included two "broad" topics that asked about a general subject (e.g. *Conference on Birth Control*) and two "narrow" topics that asked about some specific event (e.g., *Nobel Prize for Economics in 1994*). Relevance judgments for the top-50 documents for each topic were also known, but those judgments were used only to evaluate the results after the experiment was completed. As might be expected, it turned out that in every case there were more relevant documents in the top-50 for the broad topics than for the narrow ones.

The iCLEF experiment was designed in a manner similar to that used in the TREC Interactive Track, in which a Latin square design is used to block topic and searcher effects so that the system effect can be characterized. Table 1 shows the order in which topic-system combinations were presented to users. In this design, every searcher sees all four topics, two with one system and two with the other. The order in which topics and systems are presented is varied systematically in order to minimize the impact of fatigue and learning effect on the observability of the system effect. We realized at the outset that four participants was an undesirably small number given the large variability that has been observed in human performance of related tasks, but time and resource limitations precluded our use of a larger sample.

Participant	Before break		After break	
umd01	MT	Topic 11, Topic 17	Gloss	Topic 13, Topic 29
umd02	Gloss	Topic 11, Topic 17	MT	Topic 13, Topic 29
umd03	MT	Topic 17, Topic 11	Gloss	Topic 29, Topic 13
umd04	Gloss	Topic 17, Topic 11	MT	Topic 29, Topic 13

Table 1. iCLEF-2001 experiment design as run. Topics 11 and 13 are broad, Topic 17 and 29 are narrow.

The task to be performed at each participating site included:

- Design and implement two interactive document selection systems. Use of the Systran translations was optional, but we choose to use them as our MT system.

- Have participants make relevance judgments for each topic. Each participant was allowed 20 minutes for each topic (including reading the topic description, reading as many documents or document summaries as time allowed, and making relevance judgments). For each document, the participant was asked to select one of four possible judgments: "not relevant," "somewhat relevant," "relevant," or "unsure." A "not judged" response was also available.
- Ask each searcher to complete questionnaires regarding their background, each search, each system, and their subjective assessment of the two systems.
- Provide the participants judgments to the iCLEF coordinators in a standard format for scoring.
- Conduct data analysis using the scored results and other measurements that were recorded and retained locally.

An unbalanced version of van Rijsbergen's F measure was selected for use as the official effectiveness measure for the evaluation:

$$F_\alpha = \frac{1}{\alpha/P + (1-\alpha)/R}$$

where P is precision and R is recall. Values of α could range between 0 and 1, with values above 0.5 emphasizing precision and values below 0.5 emphasizing recall [5]. For iCLEF, 0.8 was selected as the value for which the experiments were to be designed, modeling a situation in which finding documents accurately is more important than finding all the relevant documents. The participants were told that they should approach the task with that in mind. For the official results, judgments of "somewhat relevant," "unsure," and "not judged" were treated as "not relevant."

3 Maryland iCLEF Experiments

For the past few years, our team at Maryland has focused on low-cost techniques for extending CLIR capabilities to new languages. Our initial work was based on using existing bilingual term lists to perform dictionary-based CLIR, and it is that technique that we adapted to perform gloss translation for these experiments. The basic idea is to find source language (in this case, French) terms in the bilingual term list and then replace them with the corresponding target language term(s) (in this case, English). For resource-poor languages we could conceivably obtain bilingual term lists by scanning (or even rekeying) a printed bilingual dictionary or by training a statistical translation model on translation-equivalent text pairs that might be automatically farmed from the Web—for these experiments we used a bilingual term that we had downloaded form the Web for CLEF 2000 [5]. This resource contained approximately 35,000 term pairs.

3.1 Gloss Translation

Bilingual term lists found on the Web often contain an eclectic combination of root and inflected forms. We therefore applied the same backoff translation strategy that we have previously used for automatic retrieval to extend the source-language (French) coverage of the term list. The first step was to remove all punctuation and convert every character to unaccented lower case in both the documents and the term list. This had the effect of minimizing problems due to character encoding. The translation process then proceeded in the normal reading order through the text, using greedy longest string matching to identify terms in the document that can be translated using the bilingual term list. If no multi-word or single word match is found, the French word in the document is stemmed and a match with the term list is attempted again. If that fails, the previous step is repeated using a second version of the bilingual term list in which all source language terms have been stemmed.[3] If the source-language term was still not found in the term list, it was copied unchanged into the translated document. We used the stemmer that we had developed for CLEF 2000 for this purpose. Bilingual term lists typically contain several possible translations for some terms. In past work, we have explored display strategies for presenting multiple alternatives, but for our iCLEF experiments we chose only a single translation for each term because we wanted to focus on a single factor (the translation strategy). As we have before, we chose the English translation that occurred most often in the Brown Corpus (a balanced corpus of English) when more than one possible translation was present in the term list.

3.2 Machine Translation

Maryland also performed full machine translation, contributing the results for use as the translations that were provided to all participating teams. Producing of the English translations of the French documents was relatively straightforward. First, we used Systran Professional 3.0 to translate the French collection into English. We then corrected some SGML tags that were inadvertently translated or mangled in some way (e.g., white spaces was inserted within the tags) and corrected them using a simple Perl script. After the translated collection was released, we found some additional mangled SGML symbols in the document titles, so we deleted these symbols. Punctuation and untranslatable words are handled differently by Systran—punctuation and upper/lower case are retained and untranslatable words are displayed in upper case with accents retained.

3.3 User Interface

Because we wished to compare translation strategies, we sought to minimize the effect of presentation differences by using the same user interface with both types

[3] Multi-word expressions in the source language are removed from the stemmed term list, so only single-word matches are possible in these last two steps.

of translation. The user interface for our experiment was based on an existing system that we had developed for our TREC-9 CLIR track experiments [4]. The system uses a Web-based server-side architecture. Searchers interact with the system using a Web browser, and their relevance judgments are recorded by the server when a search is completed. A search starts when a searcher selects a topic and a translation option (MT or Gloss) and ends when the relevance judgments for that topic are finished. A search-ID is assigned to each search so that multiple searches can be tracked simultaneously, but participants in the study completed the task individually so this capability was not needed. The system included the following capabilities:

- Provide topic selection and translation option selection mechanisms.
- Display topic descriptions based on the searcher's selection. The topic is displayed separately prior to the ranked document list so that the searcher can read and understand it before making any relevance judgments, and it remains displayed at the top of the page once the ranked list is displayed as a ready reference.
- Display a ranked list providing summary information for the top 50 documents for the selected topic (see Figure 1). The summary information that we displayed for this experiment is simply the translation of its title, as specified by the appropriate SGML tag. Query terms (i.e., any term in the topic description) that appeared in a translated summary were detected using string matching and highlighted in red and rendered in italics. A set of five radio buttons under each title allowed relevance judgments to be selected, with "not judged" initially selected for all documents.
- Display the translation of the full text of a document in a separate window whenever that document is selected by a searcher. All translations are performed in advance and cached within the server, so no speed difference between translation types is apparent to the searcher. Again, query terms that appeared in a translated document were highlighted in red and rendered in italics.
- Record the amount of time spent on judging each document. This was implemented with a Javascript timer built in a CGI script. The timer was started when the title link was selected, and stopped when one of the relevance judgment radio buttons was selected. One can easily see this method fails to record the time correctly if the judgment was based solely on a displayed summary since in that case the title link would never be selected. It would be hard to do better without an eye tracker, since multiple summaries are displayed on the same page. On the other hand, since the summaries are very short (often only one line on the screen), the time required to render a judgment in such cases is likely to be quite small.
- Simultaneously record the relevance judgments for all documents when a search is completed. This design allows users to make a quick pass through the documents and then go back for a more detailed examination if they desire. The submit button is at the bottom of the ranked list page (not shown in the Figure 1).

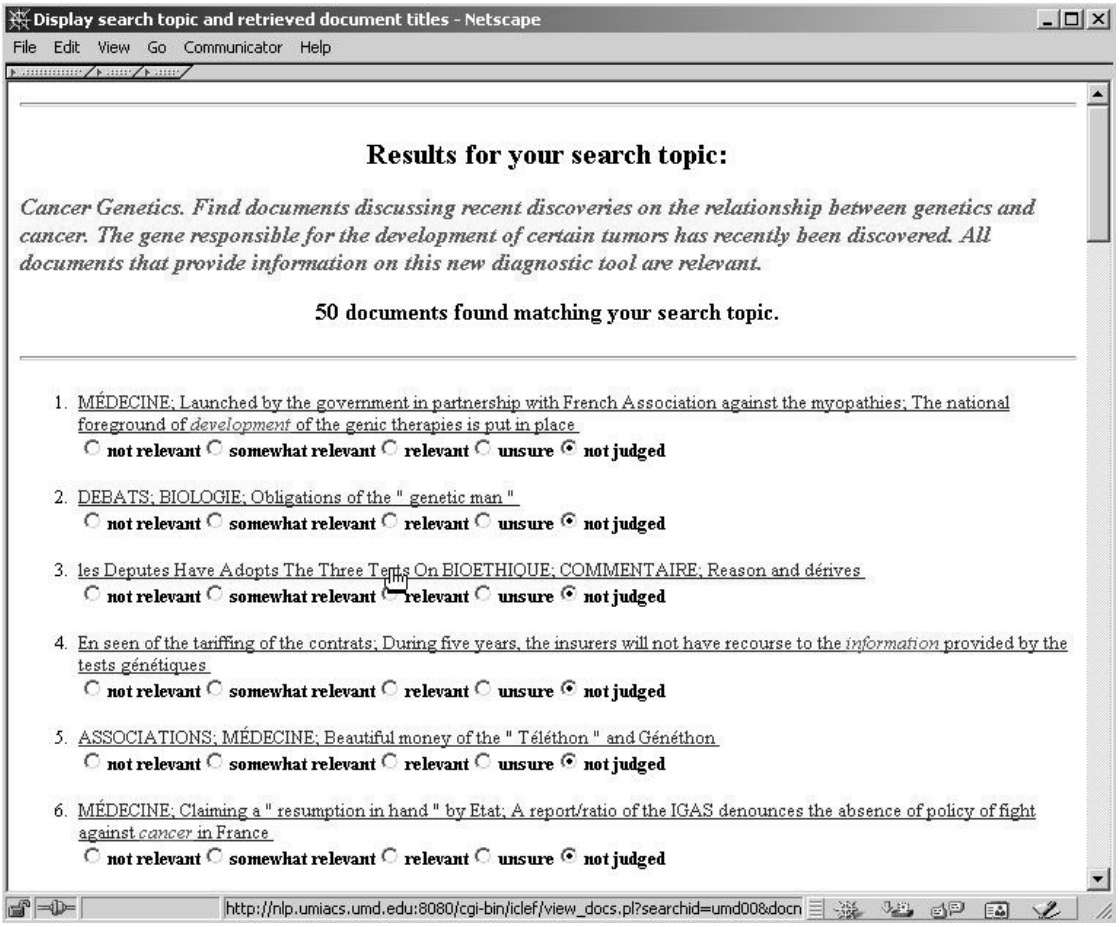


Fig. 1. Display of the ranked list of documents (MT).

3.4 Searcher Characteristics

We had originally intended to recruit graduate students in library science to participate in our experiment, since we expect that librarians could make extensive use of CLIR systems when conducting searches on behalf of people with different language skills. The fact that the experiments were performed during summer session limited the pool of potential participants, however, and the 3-hour search session made participation less appealing even though we offered a cash payment ($20) to each participant. As the deadline approach, we therefore became somewhat less selective. Of the four participants in our experiments, two (umd01 and umd03) held a Masters degree in Library Science. Both of those participants were doctoral students in the College of Information Studies, and both have interests in information retrieval and human-computer interaction.

A third subject (umd02) has a Masters degree in Computer Science and some familiarity with cross-language retrieval and is working as a user interface programmer. The fourth participants (umd04) has a Bachelors degree in religion, is currently working as a financial controller, and professed no interest in the technical details of what we were doing.

The ages of the four participants range were between 28 and 35 at the time of the experiment. None of the participants had been involved in previous interactive retrieval experiments of this sort, but all had at least five years of online searching experience. All four participants reported a great deal of experience searching the World Wide Web and a great deal of experience of using a point-click interface. Our observations during the experiment agreed with their assessments on this point.

In addition to the backgrounds described above, the following self-reported characteristics of distinguished an individual participants from the group:

umd01. Participant umd01 reported 14 years of searching experience, much more than any other participant.

umd03. Participant umd03 was the only one to report good reading skills in French (the others reporting poor skills or none). Knowledge of French was disallowed by the track guidelines, and we had mentioned this when recruiting subjects. When we saw this answer on the questionnaire at the beginning of the session, we were therefore somewhat surprised. Unfortunately, there was not sufficient time remaining before the deadline to recruit an additional participant. Interestingly, after the experiment, participant umd03 mentioned in a casual conversation that they had studied French in high school. Clearly we need to give more thought to how we conduct language skills screening.

umd04. Participant umd04 was the only one of the four with no experience searching online commercial systems, the only one to report typically searching less than once a day (for umd04, the response was twice a week) and the only one to give a neutral response to the question of how they feel about searching (the others reporting that they either enjoy or strongly enjoy searching).

3.5 Experiment Procedure

The iCLEF experiment in Maryland started on June 27, 2001, and ended on July 9, 2001. We began with a small (two-user) pilot study, after which we made some changes to our system. We then conducted a half-hour peer review session with several graduate students who were working on computational linguistics. After a few further changes, we froze the configuration of the interface for the experiments reported in this paper.

The four search sessions were conducted individually by the first author of this paper. Upon arrival, a searcher was first given a 10-minute brief introduction to the goal of the study, the procedure of the experiment, the tasks he or she was expected to complete, and the time allocation for each step. Then a 5-minute

pre-search questionnaire was completed. The major purpose of that questionnaire was to collect basic demographic information and information about the searcher's experience with searching, using point-click interface, and reading the document language. Following that was a 30-minute tutorial in which the two systems were introduced. The tutorial was conducted in a hands-on fashion—the searcher practiced using the systems while reading printed instructions line-by-line. The experimenter followed along with the searcher, pointing out specific details that might have been incompletely understood when necessary. We found that all the searchers learned how to use the systems in less than 30 minutes. After this step, the searcher was asked to take a 10-minute break. Interestingly, no participant thought this break was necessary, and none took it. The first search then started.

For each search, the experimenter would tell the participant which topic and system to select, and then the experimenter would quietly observe the search process and take observation notes. Participants did occasionally ask questions of the experimenter, but we tried to minimize this tendency. Each search was followed by a 5-minute questionnaire regarding the searcher's familiarity with the topic, the ease of getting started with making relevance judgments for that topic, and their degree of confidence in the judgments that they had made. When two searches with the same system were completed, a questionnaire regarding the searcher's experience with that system was conducted. That was followed by a 10-minute break and then the process was repeated with the second system. After all the four searches were completed, an exit questionnaire was completed. That questionnaire sought the participant's subjective comparison of the two systems and provided an unstructured space for additional comments.

4 Results

The hypotheses that we wished to test was that MT and gloss translation can both support effective interactive cross-language document selection. Formally, we seek to reject two null hypotheses:

- The F_α measure achieved by gloss translation could be achieved by following a rule that does not involve looking at the translations at all.
- The F_α measure achieved using the MT system is the same as that which would be achieved using the gloss translation system.

In this section we first examine the results using the official measure ($F_{0.8}$), then look at two variants on the computation of F_α, and then conclude by suggesting some alternative metrics that could prove to be useful in future evaluations.

4.1 Official Results

Table 2 shows the official results on a per-search basis, and Table 3 shows the result of averaging the $F_{0.8}$ measures of the two participants that experienced each condition. Three of the four searchers did better with MT than gloss translation

on broad topics, and all four searchers did better with MT on narrow topics. A two-tail paired t-test (p¡0.05), found no significant difference in either case, however, at $p < 0.05$). This is probably due to the an insufficient numbers of degrees of freedom in our test (i.e., too few participants), since the trend seems quite clear. So although we cannot reject the second of our null hypothesis, the preponderance of the evidence suggests that MT is better for this task than our present implementation of gloss translation when scored using the official measure.

	MT				GLOSS			
Searcher	Topic11	Topic13	Topic17	Topic29	Topic11	Topic13	Topic17	Topic29
umd01	0.62		1			0.28		0.78
umd02		0.34		0.78	0.13		0	
umd03	0.13		1			0.10		0
umd04		0.13		0.90	0.27		0.83	

Table 2. $F_{0.8}$ by search, as run, strict relevance (official results).

A couple of observations are easily made from Table 3. The values of $F_{0.8}$ for narrow topics are consistently higher than the values for broad topics. This suggests that searchers are typically able to make relevance judgments more accurately for narrow topics than for broad ones. Another interesting observation is that the values of $F_{0.8}$ for broad topics exhibit a strong central tendency by clustering fairly well around the mean, for narrow topics the values have a bimodal distribution with peaks near zero and one.

Topic	Broad		Narrow		Average	
Searcher	MT	GLOSS	MT	GLOSS	MT	GLOSS
umd01	0.62	0.28	1	0.78	0.81	0.53
umd02	0.34	0.13	0.78	0	0.56	0.07
umd03	0.13	0.10	1	0	0.52	0.05
umd04	0.13	0.27	0.9	0.83	0.52	0.55
Average	0.31	0.20	0.92	0.41	0.61	0.29

Table 3. Average $F_{0.8}$ by topic type and system, strict relevance (official results).

In order to test our first null hypothesis, we must construct some simple strategy that does not require looking at the documents. One way to do this is to simply selects all 50 documents in the ranked list as relevant. That guarantees

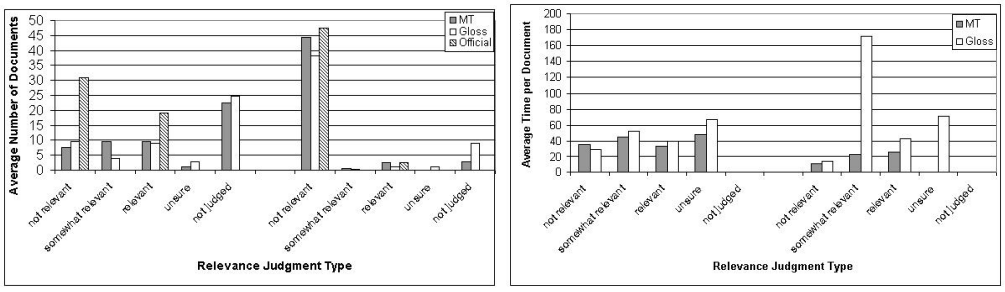
a recall of 1.0 (since we compute recall over the relevant documents in the top-50, not over all relevant documents known to CLEF). The precision is then the fraction of the entire list that happens to be relevant, which is much larger for broad topics than narrow ones. The average over all topics for $F_{0.8}$ when computed in this way is 0.26. All participants beat that value by at least a factor of two when using the MT system, and two of the four participants beat it by that much when using gloss translation. From this we tentatively conclude that both MT and gloss translation can be useful, but that there is substantial variation across the population of searchers with regard to their ability to use gloss translations for this purpose. The first part of this conclusion is tentative because we have not yet tried some other rules (e.g., always select the top 10 documents, or select different numbers of documents for broad and narrow topics) that might produce higher values for $F_{0.8}$.

4.2 Descriptive Data Analysis

No single measure can reflect every interesting aspect of the data, so we performed some descriptive data analysis to further explore our results. Figure 2 (a) shows the average number of documents to receive each type of relevance judgment by topic and system type. In that figure, we treat the official CLEF judgments as a third "system" for which only two types of judgment were provided. Clearly, many more documents were left unjudged for broad topics than for narrow ones. The highly skewed distribution of judgments on nonrelevant documents is particularly striking, suggesting that there is something about narrow topics that helps users to make more total judgments and to get the balance between relevant and not relevant judgments about right, regardless of the system type. One other observation that we could make is that for broad topics, our participants seemed to exhibit a greater proclivity to assess documents as relevant than as not relevant (based on the fraction of the official judgments that they achieved in each category). That may, however, be an artifact of the presence of a greater density of truly relevant documents near the top of any well constructed ranked list.

Examining the time required to make relevance judgments provides another perspective on our results. As Figure 2 (b) shows, "unsure" and "somewhat relevant" judgments took longer on average than "relevant" judgments, and "not relevant" judgments could be performed the most quickly. This was true for both topic types, and it helps to explain why narrow topics (which have few relevant documents) had fewer "not judged" cases. The seemingly excessive time required to reach a judgment of "somewhat relevant" when using gloss translation results from a single data point, and therefore provides little basis for any sort of inference.

The total number of documents of each relevance judgment type (across both topic types) is: "not relevant:" 398, "somewhat relevant:" 57, "relevant:" 89, and "unsure:" 20. Comparing these numbers with the average amount of time per document of each relevance type in Figure 2 (b), we see a clear inverse relationship between the number of documents and time required to assign a document



(a) (b)

Fig. 2. (a) Average number of judgments (b) Average time per judgment, by judgment type. In each chart, broad topics are on the left and narrow topics are on the right.

to that category. One possible explanation for this would be a within-topic learning effect, in which searchers learn to recognize documents in a category based on their recollection of documents that have been previously assigned to that category. Our observation of search behavior offers some evidence to support this speculation. We observed that some searchers often modified their relevance judgment, either right afterwards or later when they worked on a different document. In that second case, presumably their judgment of the relevance of the later document seemed to be related to the relevance of a previously judged document. We observed that other searchers rarely changed their relevance judgments, however, so it is not clear how pervasive this effect is.

It is interesting to note that the track guidelines did not provide any formal definition for the types of relevance judgments, presumably assuming that both experimenters and searchers would understand them based on the common meanings of the terms. In our study, we provided no further explanation of the judgment types to our participants, and no searcher expressed any confusion regarding this terminology. For this reason, we decided to explore whether the participants interpreted these terms consistently. That is the focus of the next subsection.

4.3 Comparing Strict and Loose Relevance Judgments

For the official results "somewhat relevant" was treated as "not relevant." For the sake of brevity, we will refer to that as "strict" relevance judgment. We could equally well choose to treat "somewhat relevant" as "relevant," a scenario that we call "loose" relevance judgments. Our key idea was simple: we recomputed the $F_{0.8}$ measure with all "somewhat relevant" judgments treated as "relevant," and if the measure increased, it would indicates that on average the participants were being stricter than necessary in making their relevance judgments. Table 4 shows the $F_{0.8}$ value by search with loose relevance judgments, and Table 5 compares the average $F_{0.8}$ value by systems and judgment type. Higher values

are obtained from loose judgments in both cases, but the improvement is far larger for gloss translation than for MT.

	MT				GLOSS			
Searcher	Topic11	Topic13	Topic17	Topic29	Topic11	Topic13	Topic17	Topic29
umd01	0.80		1			0.35		0.80
umd02		0.29		0.65	0.38		0	
umd03	0.74		1			0.17		0
umd04		0.20		0.67	0.68		1	

Table 4. $F_{0.8}$ by search, as run, loose relevance.

	Average $F_{0.8}$	
Relevance	MT	GLOSS
Strict	0.61	0.29
Loose	0.67	0.42
Relative improvement	10%	45%

Table 5. Comparison of strict and loose relevance.

Figure 3 depicts this difference for each of the 16 searches, with bars above the X axis indicating that loose judgments produce higher values and values below the axis indicating that strict judgments would have been better. Two trends are evident in this data. First, broad topics benefit more from loose judgments than narrow topics. Second, the improvement for gloss translation was more consistent than the improvement for MT. There were 40 judgments of "somewhat relevant" for MT, but only 17 for gloss translation, so more does not seem to be better in this case. It seems that the "somewhat relevant" judgments that people made with MT and and gloss translation were actually different in some fundamental way. One possibility is that our participants treated "somewhat relevant" as a variant of "unsure," perhaps assigning "somewhat relevant" when they had some inkling that a document might be relevant (i.e., there were not completely unsure).

4.4 Recall-Oriented Measures

It is not possible to determine how a recall-oriented searcher would have behaved from our data because we gave the searchers instructions that we expected would cause them to be biased in favor of precision. Nonetheless, we can gain some

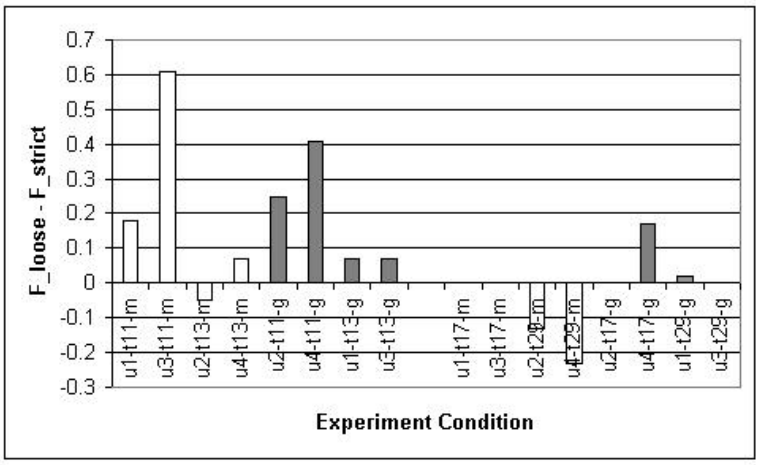


Fig. 3. Effect of loose (better above axis) and strict (better below axis) relevance on $F_{0.8}$. Left: broad topics, right: narrow topics. Each bar is labeled with searcher-topic-system (e.g., u1-t11-m means searcher umd01, Topic11, MT).

insight into the behavior of recall-oriented measures by computing $F_{0.2}$ rather than $F_{0.8}$. Table 6 shows the average values for $F_{0.2}$ by topic and system type with strict judgments. Comparison with Table 3 shows that MT and gloss translation now achieve comparable results on broad topics, with one searcher doing better with gloss, a second doing better with MT, and the other two doing poorly with both. The results for narrow topics are more consistent, with MT beating gloss translation for every searcher for with both precision-oriented and recall-oriented measures. This should not be too surprising, however, since there are so few relevant documents to be found in the case of narrow topics that recall may not be a discriminating factor.

Topic	Broad		Narrow		Average	
Searcher	MT	GLOSS	MT	GLOSS	MT	GLOSS
umd01	0.33	0.43	1	0.93	0.67	0.68
umd02	0.52	0.03	0.93	0	0.73	0.01
umd03	0.03	0.09	1	0	0.52	0.05
umd04	0.09	0.08	0.70	0.55	0.40	0.31
Average	0.24	0.16	0.91	0.37	0.57	0.62

Table 6. Average $F_{0.2}$ by topic type and system, strict relevance.

4.5 Individual Differences

Figure 4 shows the average F_α for each participant using all four variants of that measure that were defined above (two values for α, with strict and loose relevance for each). Clearly, participant umd01 outperformed the other three, regardless of what measure we use. Recall that umd01, who reported far more experience with online searching than any other participant, is now working as an IR system designer. Participant umd01 judged 186 of the 200 available documents in the time allowed, on average the other 3 participants could judged an average of 141 documents. Since most unjudged documents were for broad topics, for which an average of almost 40% of the documents were relevant, our measures penalized searchers more for failing to finish their judgments for broad than for narrow topics.

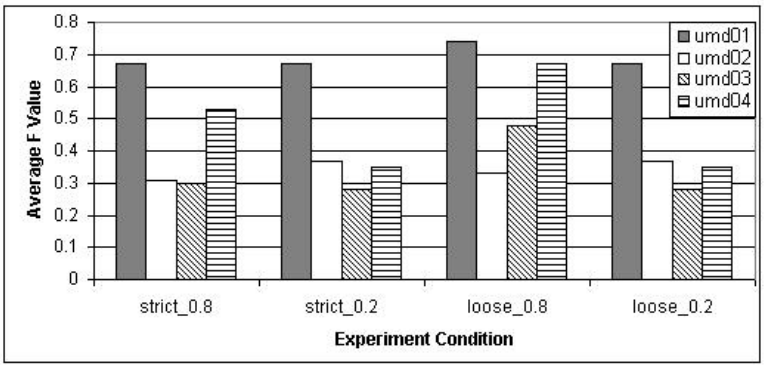


Fig. 4. Average F_α by searcher, α, and relevance type.

Two other factors that had been of potential concern to us turned out not to make much of a difference. The first of these was that participant umd03 reported that they had good reading skills in French. As Figure 4 shows, that participant actually achieved the lowest average values for three of the four measures (although two or three other participants were close in every case), and Table 3 shows that this poor performance was consistent for both MT and gloss translation. The other factor we had concern about was that some of the subjects might actually know quite a bit about one of the topics. This actually did happen in one case, again with searcher umd03, for topic 29. As it turned out, the value of F_α for that search was zero for both values of α. Go figure.

4.6 Subjective Evaluation

After each experiment, we solicited comments from our participants on the two systems and their degree of confidence in the relevance judgments that they had made. All searchers found the gloss translations were difficult to comprehend,

and three of the four participants indicated that it was difficult or very difficult to judge the relevance of documents using gloss translations. All three of those participants felt that their judgment would have been even more accurate if they had been able to look at higher quality translations. The exception was participant umd01 who thought it was easy to judge relevance with gloss translations and had confidence in the judgments they made with that system. All participants felt that it was easy to make relevance judgments with the MT system, and three of the four indicated that they liked the translation quality (umd02 didn't comment). Two felt that an even higher quality translation could still make relevance judgment much easier, while the other two thought it would only help a little bit.

In comparing the two systems, two participants felt that the difficulty of learning to use the two systems was comparable, while the other two felt that the MT system was easier to learn. Three of the four found the MT system easier to use while the remaining participant (umd01 again) found the gloss translation system easier to use. In amplifying on this, participant umd01 wrote that they believed that the gloss translation system seems easier to browse for "factual search questions."

5 Conclusion

Given that iCLEF is the first multi-site evaluation of interactive cross-language document selection, we are quite satisfied with the degree of insight that our experiments have provided. Our results suggest that both full machine translation and simple term-for-term gloss translation strategies provide a useful basis for selecting documents in an unfamiliar language, but that there is substantial room for improvement over our present gloss translation technique for this task. Perhaps more importantly, we have found insight in our data into factors that we had not previously considered, such as the importance of providing clear facilities for distinguishing between uncertainty and partial relevance. We have also learned something about the strengths and weakness of our present measures, with perhaps the most important point being that narrow topics pose a fundamentally different search task than broad topics. Perhaps we will ultimately find that it would be best to model those different tasks using different effectiveness measures. This first iCLEF has indeed pointed the way towards an interesting and important set of questions, but much remains to be done.

Acknowledgments

The authors would like to thank Clara Cabezas for assistance in setting up the systems used in the study, Gina Levow for help with gloss translation, Bob Allen for advice on statistical significance testing, our participants for their willingness to invest their time in this study, and members of the CLIP lab for performing peer review of our system. This work has been supported in part by DARPA cooperative agreement N660010028910.

References

1. William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kraemer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluations give the same results? In *Proceedings of the 23nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 17–24, August 1998.
2. Douglas W. Oard. Evaluating interactive cross-language information retrieval: Document selection. In Carol Peters, editor, *Proceedings of the First Cross-Language Evaluation Forum*. 2001. http://www.glue.umd.edu/∼oard/research.html.
3. Douglas W. Oard and Anne R. Diekema. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, volume 33. American Society for Information Science, 1998.
4. Douglas W. Oard, Gina-Anne Levow, and Clara I. Cabezas. TREC-9 experiments at Maryland: Interactive CLIR. In *The Ninth Text Retrieval Conference (TREC-9)*, November 2000. http://trec.nist.gov.
5. Douglas W. Oard, Gina-Anne Levow, and Clara I. Cabezas. CLEF experiments at Maryland: Statistical stemming and backoff translation. In Carol Peters, editor, *Proceedings of the First Cross-Language Evaluation Forum*. 2001. http://www.glue.umd.edu/∼oard/research.html.