The TREC-2002 Arabic/English CLIR Track

Douglas W. Oard
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
oard@glue.umd.edu
and
Fredric C. Gey
UC Data Archive & Technical Assistance
University of California, Berkeley, CA 94720
gey@ucdata.berkeley.edu

Abstract

Nine teams participated in the TREC-2002 cross-language information retrieval track, which focused on retrieving Arabic language documents based on 50 topics that were originally prepared in English. Arabic translations of the topic descriptions were also made available to facilitate monolingual Arabic runs. This was the second year in which a large Arabic document collection was available. Three new teams joined the evaluation, and the cross-language aspect of the evaluation received more attention this year than in TREC-2001. A set of standard linguistic resources was made available to facilitate cross-system comparisons, and their use as a contrastive condition was encouraged. Unique contributions to the relevance pools were more typical of previous TREC evaluations then the results of TREC-2001 had been for the same document collection, with no run uniquely contributing more than 6% of the known relevant documents.

1 Introduction

The goal of the 2002 Text Retrieval Conference (TREC-2002) Cross-Language Information Retrieval (CLIR) task was to develop evaluation methodologies and evaluation resources to assess the effectiveness of ranked retrieval techniques that accept English queries and search Arabic documents. Monolingual Arabic experiments, in which both the queries and the documents were in Arabic, were also supported. Standard French translations were also provided in TREC-2001, but that was not done this year because no team expressed interest in working with French in preference to English. TREC-2002 was the ninth year in which non-English document retrieval has been evaluated at TREC, and the sixth year in which cross-language information retrieval has been the principal focus of that work. For a summary of prior evaluations, readers are referred to [1]. In this paper we describe the task and the evaluation collection, we summarize the techniques used by each participating team and briefly describe their results, and we offer some guidelines for future use of the TREC Arabic collection.

2 Task Description

As in past CLIR evaluations, the principal task for each group was to automatically build queries from topic descriptions written in one language (English, in this case) and then use those queries as a basis for ranking documents written in another language (Arabic, in this case) in order of decreasing degree (or probability) of topical relevance. Each participating team was allowed to submit as many as five runs for official scoring. In order to foster comparability, teams submitting cross-language runs were required to submit at least one run in which only the title and description fields of the topic description were used. Evaluation then proceeded by pooling the top 100 documents for each topic from each of the 41 submitted runs, manual examination each document in the pool by a human judge (usually the creator of the topic), and recording

binary (yes/no) topical relevance judgments for each document in each topic's judgment pool. Participating teams were also invited to perform additional "post-hoc" runs, scoring them locally using relevance judgments provided by NIST, if they wished to investigate more conditions than would be possible using only the five official runs. The top 1000 documents in the ranked list for each topic was evaluated using a suite of measures. In this paper, we report the mean (over 50 topics) of the precision at 11 levels of recall and the mean (again, over 50 topics) of the uninterpolated average precision. Additional statistics for each run can be found elsewhere in this proceedings.

2.1 Topics

For TREC-2002, fifty topic descriptions (numbered AR26-AR75) were created in English in a collaborative process between the LDC and NIST (for TREC-2001, only 25 topics had been created). An example from this year's topic set is:

```
<top>
<num>Number: AR26</num>
<title>Kurdistan Independence</title>
<desc> Description:
How does the National Council of Resistance relate to the potential
independence of Kurdistan? </desc>

<narr> Narrative:

Articles reporting activities of the National Council of Resistance
are considered on topic. Articles discussing Ocalan's leadership
within the context of the Kurdish efforts toward independence are
also considered on topic.</narr>

</top>
```

The Linguistic Data Consortium also prepared an Arabic translation of the topics, so participating teams also had the option of doing monolingual (Arabic-Arabic) retrieval. The same topic in Arabic was distributed as:

<top>
<num> Number: AR26 </num>
<title> مجلس المقاومة الوطني الكردستاني </title>
<desc> Description:

كيف ينظر مجلس المقاومة الوطنية الى الإستقلال المحتمل للاكراد؟

</desc>
<narr> Narrative:
الموضوع يتضمن نصوص متعلقة بتحركات مجلس المقاومة الوطنية ، مقالات تتحدث عن قيادة اوجلان ضمن جهود الاكراد للاستقلال .

</narr>
</top>

2.2 Documents

As in the TREC-2001 CLIR track, the document collection contained 383,872 newswire stories that appeared on the Agence Française de Presse (AFP) Arabic Newswire between 1994 and 2000. The documents were represented in Unicode and encoded in UTF-8, resulting in an 896 MB collection. A typical document is shown in Figure 1.

<DOC>
<DOCNO>20000321_AFP_ARB.0001</DOCNO>
<HEADER>ارا0100 4 ش 8920 قبر /اف-ذ ز ز03 اسرائيل/فلسطينيون</HEADER>
<BODY>
<HEADLINE>جرح ثلاثة اسرائيليين اصابة اثنين منهم خطيرة في هجوم في الضفة الغربية</HEADLINE>
<TEXT>
<P>القدس 12-3 (اف ب) - افادت حصيلة جديدة للجيش الاسرائيلي ان ثلاثة اسرائيليين جرحوا مساء امس الاثنين في هجوم جرى عندما اطلق عليهم الرصاص من سيارة تجاوزت السيارة المدنية التي كانت تقلهم قرب ترقومية في محيط الخليل بالضفة الغربية.</P>
<P>واوضح المتحدث باسم الجيش الاسرائيلي ان سائق السيارة التي كانت تقل الاسرائيليين، وهو من مستوطني الضفة الغربية اصيب بجروح "طفيفة". ووصف حالة احد الجريحين الاخرين بانها "حرجة" وحالة الثاني بانها "خطيرة".</P>
<P>وتشكل الخليل حيث يقيم 004 مستوطن يهودي بحماية الجيش الاسرائيلي وسط 021 الف فلسطيني، بؤرة توتر بين الاسرائيليين والعرب. وقد انسحبت اسرائيل في كانون الثاني/يناير 7991 من 08% من هذه المدينة وابقت على وجود عسكري كبير في الحي الذي يسكنه المستوطنون.</P>
<P>وجرح الاسرائيليون الثلاثة عندما تعرضت السيارة التي كانوا فيها لاطلاق نار من سيارة اخرى تجاوزتها قرب بلدة ترقومية التي يؤدي اليها "الممر الامن" الذي يربط بين غزة وجنوب الضفة الغربية مرورا بالاراضي الاسرائيلية.</P>
<P>وقد نقل الجريحان بسيارة اسعاف ثم بمروحية الى مستشفى حداسا في القدس.</P>
<P>وبدأ الجيش عمليات بحث عن الفاعلين واقام حواجز على الطرقات.</P>
<P>وابلغت السلطة الفلسطينية بملابسات الهجوم لتحاول العثور على مرتكبيه.</P>
<P>واشاد المسؤولو الاسرائيليون في الفترة الاخيرة بالتعاون مع اجهزة الامن الفلسطينية في اطار مكافحة الارهاب.</P>
<P>ونشرت بلدية مستوطنة كريات اربع القريبة من الخليل بيان احتجاج على سياسة السلام التي يبنتها رئيس الوزراء الاسرائيلي ايهود باراك الذي تتهمه "بترك المستوطنين رهائن بايدي الفلسطينيين".</P>
<P>وقال الجيش الاسرائيلي في تقديرات اولية ان خلية تابعة لحركة المقاومة الاسلامية (حماس) قد تكون وراء الاعتداء.</P>
<P>وتعارض حركة حماس بشدة اتفاقات اوسلو حول الحكم الذاتي الفلسطيني المبرمة عام 3991 وقد اعلنت مسؤوليتها عن عالبية الاعتداءات التي استهدفت اسرائيل منذ ذلك الحين.</P>
</TEXT>
<FOOTER>اف004 موا/ا شف _____</FOOTER>
</BODY>
<TRAILER>جمت مار 00 405012</TRAILER>
</DOC>

Figure 1. A sample Arabic document from the AFP collection.

3 Relevance Judgments

The nine participating teams shown in Table 1 together produced 23 automatic cross-language runs, 17 automatic monolingual runs with Arabic queries, and one manual monolingual run. The total number of relevant documents found over 50 topics was 5,909 with a mean of 118 relevant documents per topic, a maximum of 523, and minimum of 10. In TREC-2001, such a large number of relevant documents were uniquely found by individual runs that there was some concern about the comparability of post-hoc and official runs. In the pooled relevance judgment methodology, most documents remain unjudged, and the usual procedure is to treat unjudged documents as if they are not relevant. Voorhees has shown that the preference order between automatic runs in the TREC ad hoc retrieval task would rarely be reversed by the addition of missing judgments, and that the relative reduction in mean uninterpolated average precision that would result from removing "uniques" (relevant documents found by only a single system) from the judgment pools is typically less than 5% [2]. In TREC-2001 CLIR collection, 9 of 28 judged automatic runs experienced a relative reduction in mean uninterpolated average precision exceeding 10% relative when "uniques" contributed by that run were removed from the judgment pool. As Figure 2 shows, for the TREC-2002 CLIR collection, only one of 41 judged runs would experience more than 5% reduction. Figure 3 shows the unique relevant documents contributed to the final relevance pool by each group; no team uniquely contributed more than 6% of the known relevant documents (with the largest contribution coming from the one team that submitted a manual run). These results suggest that post-hoc use of the TREC-2002 collection will result in meaningful comparisons with the set of judged runs reported in this proceedings.

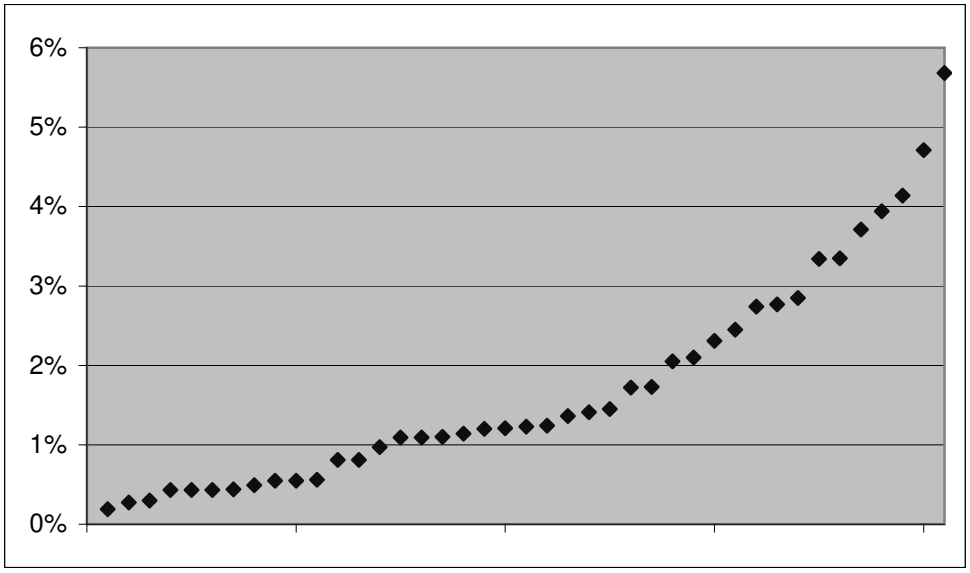

Figure 2. Effect of removing relevant documents found uniquely by each of 41 official runs.

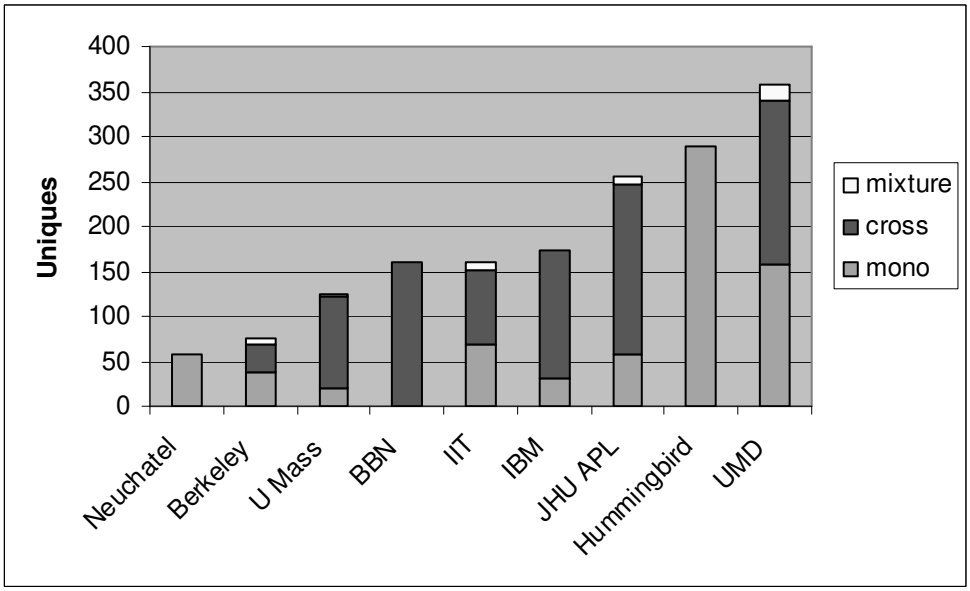

Figure 3. Unique contributions to known relevant documents, by participating team.

4 Standard Resources

Cross-language retrieval effectiveness depends on both the design of the retrieval system and the quality of the linguistic resources that are used. In order to begin to tease apart these factors, each participating team that performed the CLIR task was invited to submit one title+description run in which standard linguistic resources were used to the extent practical given their design. For example, a team using dictionary-based query translation could submit one run in which a standard bilingual dictionary was the only dictionary used. The following standard resources were made available:

- An Arabic "light" stemmer that used truncation rules to removed a small set of prefixes and suffixes. The light stemmer was developed through collaboration between Kareem

Darwish at the University of Maryland and Leah Larkey at the University of Massachusetts.

- A bidirectional Arabic-English bilingual dictionary, rekeyed from Salmone's Advanced Learner's Arabic-English Dictionary. The dictionary was provided by David Smith, of Tufts University.
- Two tables of translation probabilities, one for English-to-Arabic and the other for Arabic-to-English. These tables were developed using the Giza++ implementation of IBM Model 1 to align Arabic stems produced the track's standard light stemmer with English stems produced using the Porter stemmer. The documents on which this alignment was performed were obtained from the United Nations by Jinxi Xu at BBN and the alignments were produced by Alex Fraser of USC-ISI while working at BBN.
- A Web-based bidirectional Arabic-English machine translation system. We designated the system available at http://tarjim.ajeeb.com as the standard resource for the track.

Comparing the standard resource runs with the best runs by site (for the same query length) suggests that these resources were generally near, but not at, the state-of-the-art.

5 Retrieval Approaches

The nine groups have written papers about their methods and experiments. Three themes emerge across the reported work: (1) A greater focus on exploring innovative CLIR techniques than was evident in TREC-2001, (2) continued investigation of Arabic-specific issues, such as stemming and stopword removal, and (3) increasing reliance on multiple sources of evidence to overcome the limitation of any single source.

5.1 BBN

BBN made use of its probabilistic translation and retrieval model used in TREC-9 (for Chinese) and TREC-2002 (for Arabic) as well as the United Nations corpus. They employed the method of Hiemstra and de Jong to compute English probabilities by projecting Arabic terms to English (a weighted sum of corpus probabilities of Arabic terms where the weights are translation probabilities). To process the UN corpus and generate a new bilingual translation lexicon, BBN utilized the IBM model 4 statistical translation approach, rather than IBM model 1. For stemming they made use of the University of Massachusetts (Light8) stemmer in addition to the Buckwalter stemmer used in TREC-2001. In the area of query expansion they performed English and Arabic query expansions independently rather than sequentially (English query expansion followed by Arabic query expansion) as in TREC-2001.

5.2 University of California at Berkeley

The main focus of Berkeley's work was to develop several approaches to Arabic stemming and stopword list generation. Berkeley used the Ajeeb machine translation system to translate every word in the AFP collection after minimal word normalization. A 3,447-entry Arabic stopword list was created as the set of all Arabic terms that translated to an English stopword. A similar approach was used to generate a stemmer – Arabic words were partitioned into clusters based on their English translations, with Arabic words whose English translations were conflated to the same English stem forming a cluster. A second stemmer in which common one, two, and three character prefixes and suffixes (identified in the AFP corpus) were removed was also tried. These resources were then used in various combinations to perform dictionary-based retrieval using an extension of the logistic regression technique from previous TREC evaluations that incorporated blind relevance feedback. Berkeley also submitted a merged run in which two machine translation systems (Ajeeb and Al-Misbar) were used to perform query translation directly and the Salmone dictionary was used to perform dictionary-based query translation. Each was Score-based merging was then used to produce the final ranking.

5.3 Hummingbird Technologies

Hummingbird Technologies chose to focus on monolingual Arabic retrieval again this year. Hummingbird used a minimalist approach, with the same set of stemming rules as last year. Hummingbird makes a unique contribution to the CLIR track by evaluating commercially available technology that has been integrated into a comprehensive document management system.

5.4 Illinois Institute of Technology

In their TREC-2002 experiments with Arabic CLIR, Illinois Institute of Technology (IIT) continued their investigations of improvement of monolingual Arabic retrieval using different stemming approaches. For TREC-2001 IIT developed a 'light stemming' approach that performed well. For TREC-2002, they developed two new stemmers, one rule-based and one based on pattern matching. Both stemmers used an external training corpus from 1999-2001 editions of two Saudi Arabian newspapers to identify frequent prefixes and suffixes. Both stemmers performed comparably in monolingual Arabic retrieval, and the paper contains detailed examples of how each approach to stemming can affect word meaning. For CLIR, IIT used the Ajeeb machine translation package to translate the English queries into Arabic, and they also tried a second approach based on the translation probability tables provided as part of the standard resource set. In this case, MT produced better results, perhaps because the IIT stemmers differed from those used to produce the translation probability tables.

5.5 IBM Research

IBM's cross-language experiments were based entirely upon statistical machine translation using the IBM model 1 approach. IBM used the UN parallel corpus provided by BBN with a newly developed Arabic morphological analyzer. Two statistical machine translation systems were built. The first, an Arabic-to-English sentence translation model, was used to translate the documents into English, followed by monolingual English retrieval. The second, a probabilistic convolution model in which the probability of generating an English query stem was modeled based on the probabilities of generating that stem from an Arabic word or morpheme observed in the document. The convolution model substantially outperformed the sentence translation model, perhaps because it made use of document-wide translation probabilities.

5.6 Johns Hopkins University – Applied Physics Laboratory (JHU-APL)

JHU-APL continued its investigation of the use of overlapping character n-grams for monolingual Arabic retrieval. In TREC-2001 they used 4-grams; for TREC-2002 they investigated the use of 3-grams, 4-grams and 5-grams, and their official monolingual runs used combinations of those n-gram lengths. They used the same two machine translation systems as the Berkeley team, performing query translation from English to Arabic. The use of stemming in the standard translation probability tables made that resource unsuitable for straightforward use with n-gram-based techniques. They did, however, try mapping all English words that share a common stem to all Arabic words that share a common stem.

5.7 University of Maryland

The main focus of this year's experiments at Maryland was on combination-of-evidence techniques for cross-language retrieval. Translation knowledge was obtained from the same two machine translation systems that Berkeley used, the Salmone bilingual dictionary, and both directions of the BBN translation probability tables. Translation probabilities were estimated based on all of this evidence and then used with five CLIR techniques, one of which was a previously developed baseline (Pirkola's method). Some side experiments were also done to investigate the potential of document expansion and variants of light stemming.

5.8 University of Massachusetts

University of Massachusetts tried the most extensive set of techniques. Generally, acronyms found in the query were expanded using the U Mass *Acrophile* system. A bilingual lexicon built from a proper name dictionary from New Mexico State University and the same two MT systems that Berkeley used was then used to translate expanded English query terms and add them to the lexicon. Query expansion was performed both before and after translation. Arabic stopword removal was performed after morphological normalization using a 168-term stopword list from University of Lancaster, and several alternative forms of morphological normalization were tried. A number of variants on the processing path and the retrieval model were tried, and score-based merging among these variants was then used to create the final submissions.

5.9 University of Neuchatel

University of Neuchatel only submitted monolingual Arabic runs to the TREC-2002 CLIR Track evaluation. Neuchatel took the interesting approach of independently indexing Arabic words and of indexing tri-grams as alternative indexing and retrieval scheme, following the approach of Darwish and Oard in their SIGIR-2002 paper. Prior to all indexing and retrieval, Neuchatel converted and normalized the Arabic document text into Latin letters ("In Malta, the Arabic language is written using the Latin alphabet"). The Neuchatel word approach developed two stemmers which fall into the same area of 'light stemming' of the standard resource. They then dropped all stopwords which appeared in their 347 Arabic stopword list. From the paper it appears that their tri-gram approach utilized words as a basis (rather than including white space as other n-gram approaches usually do) and removed frequent tri-grams from a tri-gram stoplist generated from the stopword list above. Following independent retrieval of word-based indexed documents and tri-gram-based indexed documents, a "data fusion" summing approach to independently generated RSV rankings was applied to generate the final ranked list. They utilized the Rocchio approach to blind feedback and performed multiple relevance-expansion experiments on both stems and tri-grams.

6 Results

Monolingual runs establish a useful baseline to which cross-language results can be compared, and they help to enrich the relevance judgment pools. No standard query length was required for monolingual runs, but seven of the eight teams submitting monolingual runs elected to use title+description queries for at least one monolingual run. Figure 4 shows the best title+description monolingual run for those seven teams. The best of these results is somewhat below the best that was achieved last year. Since the document collection is the same, this suggests that the topics this year may be somewhat harder on average than last year's topics were.

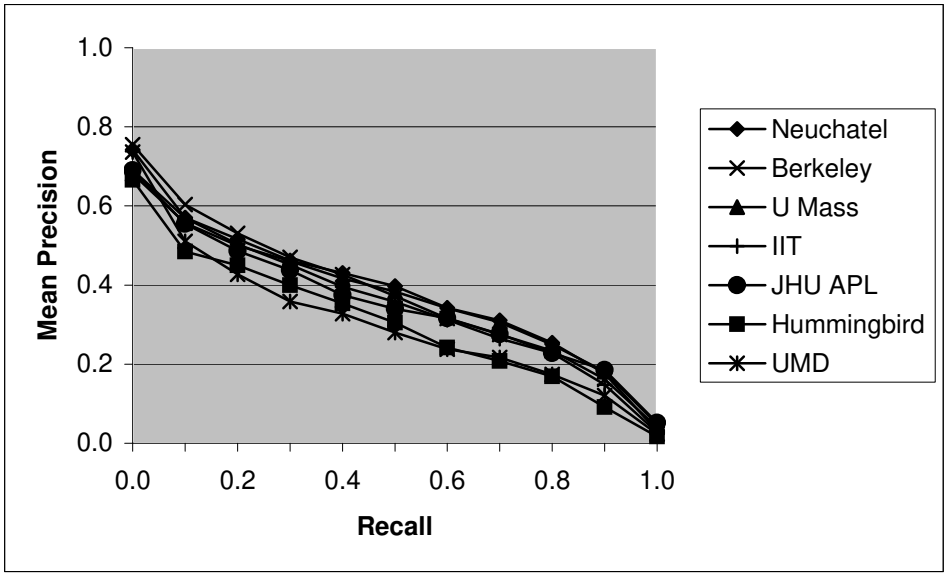

Figure 4. Best automatic monolingual runs, title+description queries.

Figure 5 shows the best automatic cross-language run for the required title+description condition for the seven teams that submitted at least one cross-language run. Five teams ran title+description queries for both the monolingual and the cross-language conditions, with the cross-language retrieval exhibiting a relative effectiveness ranging from 29% relative below monolingual to 6% relative above monolingual. It is difficult to draw any conclusions in the face of such large variability; interpretation of these results will require close attention to the implementation details of individual systems.

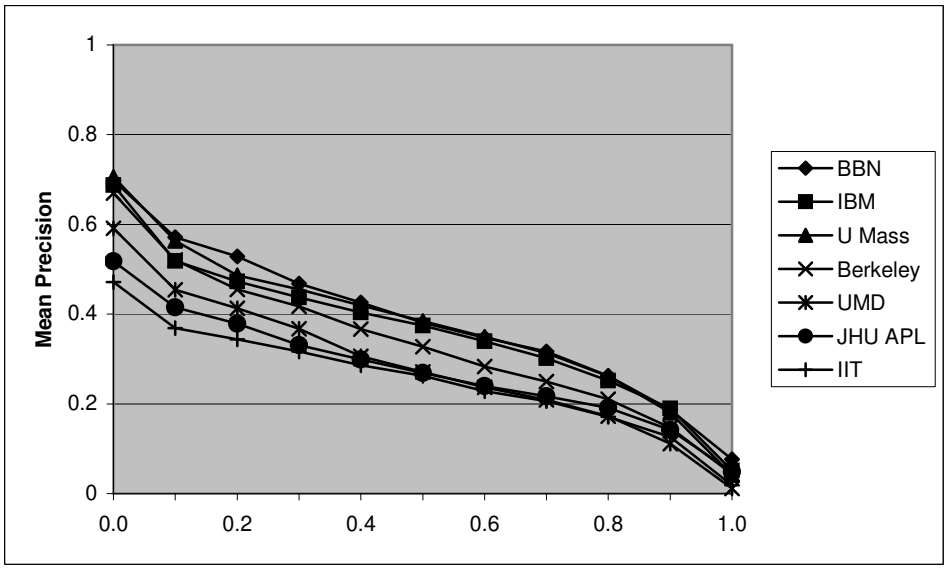

Figure 5. Best automatic cross-language runs, title+description queries.

Five teams submitted standard resource runs. As Figure 6 shows, three teams demonstrated improved performance (ranging from 4% to 11% relative) through the use of additional linguistic resources. Only five runs could be scored officially for each participating team, so post-hoc analysis may reveal greater potential for improvement from the use of additional linguistic resources than can be seen in the official results.

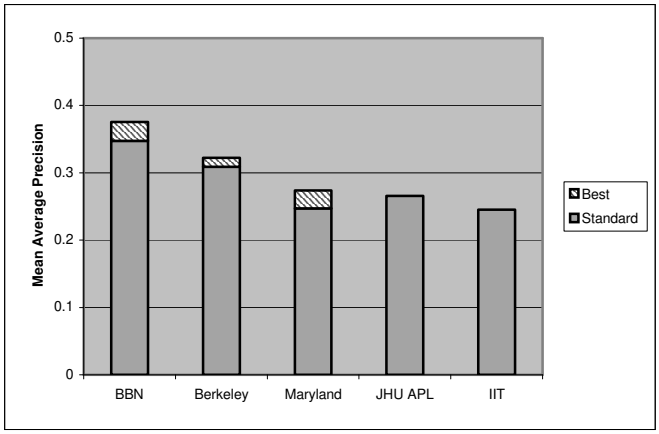

Figure 6. Improvement obtained using additional linguistic resources.

As is common in information retrieval evaluations, substantial variation was observed in retrieval effectiveness on a topic-by-topic basis. Figure 7, which plots the median and maximum average precision over the 23 cross-language runs illustrates this (note, however, that this plot includes queries of different lengths). For example, half of the runs did poorly on topics 34 and 55, but at least one run exceeded the median average precision for those topics by 800% relative. Applying this type of analysis on a run-by-run basis can help to identify effects that would be masked by averages across topics.

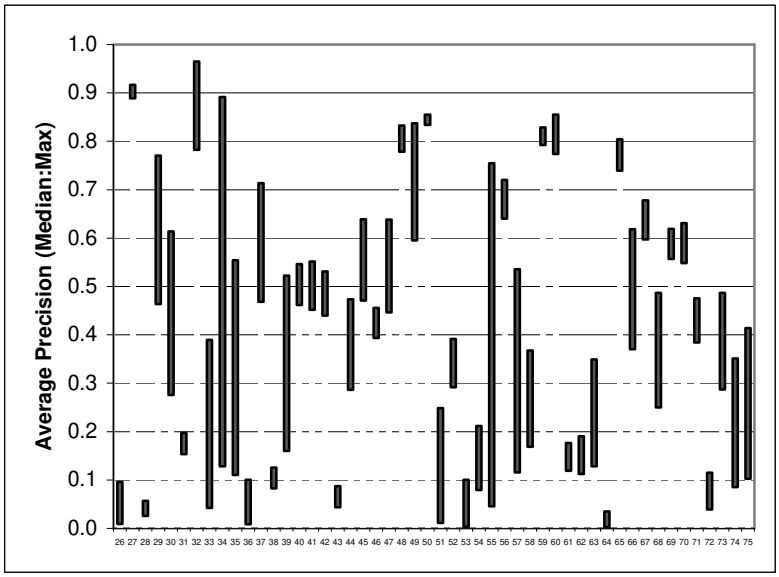

Figure 7. Average precision by topic (bottom=median, top=maximum), cross-language.

6 Looking to the Future

The TREC evaluations produce three things of enduring value: (1) research results, (2) standard test collections on which new techniques can be evaluated, and (3) a research community with shared interests. Over the past nine years, TREC has produced eight non-English test collections in six languages (Arabic, Chinese, French, German, Italian, and Spanish), and the demonstrated utility of these collections has inspired the creation of similar collections for many other languages (including Dutch, Japanese, Korean and Finnish). The results obtained this year indicate that topics AR26-AR75 are suitable for post-hoc use of the collection by automatic

systems that did not contribute to the relevance pools. Topics AR1-AR25 have proven to be of some use for system tuning, but the relatively long title fields and the markedly elevated "uniques" effect make use of that collection for comparative studies less advisable. We therefore recommend that researchers working with this collection in the future report results for the 50 topics developed this year rather than treating all 75 topics as a single collection.

Our community now includes hundreds of researchers working on CLIR in dozens of countries, and research results regularly appear in a broad array of venues. Although this is the last year of the CLIR track at TREC, similar evaluations will continue in Europe at CLEF [3] and Japan at NTCIR [4], and work on searching Arabic will continue in the Topic Detection and Tracking evaluations (TDT) [5]. The idea of providing standard resources was first tried at TDT, and we have found it to be useful in a more traditional CLIR evaluation design as well.

Some of the questions that we have explored in the CLIR track may ultimately migrate into other tracks at TREC. CLIR is, after all, just one capability among the many that are needed to build effective systems to access globally distributed information. Perhaps the future will see the creation of a multilingual Web track, a track for searching multilingual speech, or an interactive multilingual track. When that happens, the baseline technology from which those specialized applications will be built will have been first developed here, in the TREC CLIR track.

Acknowledgments

The authors are grateful to Ellen Voorhees for handling all of the logistics for this track at NIST and for providing the data that was the basis for our analysis, to the LDC for their work on topic development and relevance assessment, to the contributors of the standard resources for making that comparison possible, and to the participating research teams for their advice and insights along the way. This work was supported in part by DARPA cooperative agreements N660010028910 and N660010018911.

References

[1] Fredric C. Gey and Douglas W. Oard. The TREC-2001 cross-language information retrieval track. In: *Proceedings of the 2001 Text Retrieval Conference*, NIST, 2001.

[2] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In C.J. Van Rijsbergen W. Bruce Croft, Alistair Moffat, editor, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323. ACM Press, August 1998.

[3] Carol Peters, Martin Braschler, Julio Gonzalo and Michael Kluck, editors. Evaluation of 002 Cross-Language System Information systems: Third Workshop of the Cross-Language Evaluation Forum, CLEF-2002, Rome, Italy, Springer Computer Science Lecture Notes, forthcoming.

[4] Keizo Oyama, Emi Ishida and Noriko Kando, editors. NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research Information Retrieval, Automatic Text Summarization and Question Answering (September2001-October2002), National Institute of Informatics, Tokyo, January, 2003.

[5] James Allen, ed., Topic Detection and Tracking: Event-based Information Organization, Kluwer Academic Publishers, Boston, 2002.