The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries

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

Abstract

Ten groups participated in the TREC-2001 cross-language information retrieval track, which focussed on retrieving Arabic language documents based on 25 queries that were originally prepared in English. French and Arabic translations of the queries were also available. This was the first year in which a large Arabic test collection was available, so a variety of approaches were tried and a rich set of experiments performed using resources such as machine translation, parallel corpora, several approaches to stemming and/or morphology, and both pre-translation and post-translation blind relevance feedback. On average, forty percent of the relevant documents discovered by a participating team were found by no other team, a higher rate than normally observed at TREC. This raises some concern that the relevance judgment pools may be less complete than has historically been the case.

1 Introduction

For the 2001 Text Retrieval Conference (TREC-2001), the Cross-Language Information Retrieval (CLIR) task was to utilize English (or French) queries against Arabic documents. Monolingual Arabic experiment designs in which both the queries and the documents were in Arabic were also supported. This was the eighth year in which non-English document retrieval has been evaluated at TREC, and the fifth year in which cross-language information retrieval has been the principal focus of that work. In TREC-3, retrieval of 25 topics against a Mexican newspaper corpus was tested by four groups. Spanish language retrieval was evaluated in TREC-3, TREC-4 (another 25 topics for the same Mexican corpus), and TREC-5 (where an European Spanish corpus was used). In TREC-5, a Chinese language track was introduced using both newspaper (People's Daily) and newswire (Xinhua) sources from People's Republic of China, and 25 Chinese topics with an English translation supplied. The TREC-5 corpus was represented with the GB character set of simplified Chinese. The Chinese monolingual experiments on this collection that were done in TREC-5 and TREC-6 sparked research into the application of Chinese text segmentation to information retrieval using dictionary-based methods and statistical techniques, and simpler overlapping bigram segmentation methods were also found to be effective. TREC-6, TREC-7 and TREC-8 had the first cross language tracks, which focussed upon European languages (English,

French, German, and later Italian). Following TREC-8, the venue for European-language retrieval evaluation moved to Europe with the creation of the Cross-Language Evaluation Forum (CLEF), first held in Lisbon in September 2000 [1]. For TREC-9, the CLIR task used Chinese documents from Hong Kong. In distinction from the earlier TREC-5/6 Chinese corpus, these sources were written in the traditional Chinese character set and encoded in BIG5. Following TREC-9 the evaluation of English-Chinese retrieval moved to the NTCIR Evaluation that is coordinated by the National Institute of Informatics in Japan (http://research.nii.ac.jp/ntcir/workshop/work-en.html).

2 Task Description

As in past TREC CLIR evaluations, the principal task for each group was to match topics in one language (English or French, in this case) with documents in another language (Arabic) and return a ranked list of the top 1000 documents associated with each topic. Participating groups were allowed to submit as many as five runs, with at least one using only the title and description field of the topic description. Evaluation then proceeded by pooling ranks and manual examination of the pools by human judges to decide binary (yes/no) relevance for each document in the pool with respect to each topic. A suite of statistics were then calculated, with the mean (over 25 topics) uninterpolated average being the most commonly reported.

2.1 Topics

Twenty-five topic descriptions (numbered AR1-AR25) were created in English in a collaborative process between the LDC and NIST. An example of a topic description is:

<top>
<num> Number: AR22
<title> Local newspapers and the new press law in Jordan
<desc> Description:
Has the Jordanian government closed down any local newspapers due
to the new press law?
<narr> Narrative:
Any articles about the press law in Jordan and its effect on the local
newspapers and the reaction of the public and journalists toward the new
press law are relevant. The articles that deal with the personal suffering
of the journalists are irrelevant.
</top>

Through the efforts of Edouard Geoffrois of the French Ministry of Defense, the English topics were translated into French and made available to participants which wished to test French to Arabic retrieval. The French version of the topic shown above is:

<top>
<num> Number: AR22
<title> Les journaux locaux et la nouvelle loi sur la presse en Jordanie
<desc> Description:
Le gouvernement jordanien a-t-il interdit un journal local à cause de la nouvelle loi sur la

```
<DOC>
  <DOCNO>20000321_AFP_ARB.0001</DOCNO>
  <HEADER>ارا0100 4 ش 8920 قبر /اف-ذزز03 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 يتبعها 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 غالبية الاعتداءات
        التي استهدفت اسرائيل منذ ذلك الحين.</P>
    </TEXT>
    <FOOTER>شف/اا موا004 اف ____________</FOOTER>
  </BODY>
  <TRAILER>جمت مار 00 405012</TRAILER>
</DOC>
```


Figure 1: Example Arabic document.

presse?
<narr> Narrative:
Tout article concernant la loi sur la presse en Jordanie et ses effets sur les journaux locaux ainsi que la réaction du public et des journalistes à la nouvelle loi sur la presse est pertinent. Les articles traitant des souffrances personnelles des journalistes ne sont pas pertinents. </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.

2.2 Documents

The document collection used in the TREC-2001 CLIR track consisted of 383,872 newswire stories that appeared on the Agence France Press (AFP) Arabic Newswire between 1994 and 2000. The documents were represented in Unicode and encoded in UTF-8, resulting in a 896 MB collection. A typical document is shown in Figure 1.

3 Relevance Judgments

The ten participating research teams shown in Table 1 together produced 24 automatic cross-language runs with English queries, 3 automatic cross-language runs with French queries, 19 automatic monolingual runs with Arabic queries, and 2 manual runs (one with English queries and one with Arabic queries). From these, 3 runs were selected from each team in a preference order recommended by the participants for use in forming assessment pools. The resulting pools were formed from 15 cross-language runs with English queries, 1 cross-language run with French queries, and 14 monolingual runs with Arabic queries. The top-ranked 70 documents for a topic in each of the 30 ranked lists were added to the judgment pool for that topic, duplicates were removed, and the documents then sorted in a canonical order designed to prevent the human judge from inferring the rank assigned to a document by any system. Each document

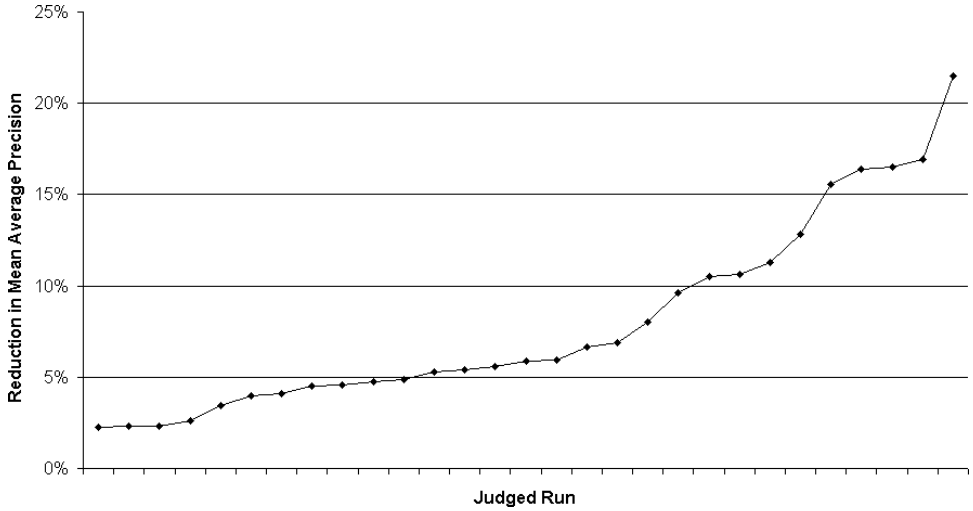


Figure 2: Effect on 29 judged runs of removing "uniques" contributed by that run.

in the pool was then judged for topical relevance, usually by the person that had originally written the topic statement. The mean number of relevant documents that were found for a topic was 165.

Most documents remain unjudged when pooled relevance assessments are used, 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 was typically less than 5% [2]. As Figure 2 shows, this effect is substantially larger in the TREC-2001 Arabic collection, with 9 of the 28 judged automatic runs experiencing a relative reduction in mean uninterpolated average precision of over 10% relative when the "uniques" contributed by that run were removed from the judgment pool.

Figure 3 helps to explain this unexpected condition, illustrating that many relevant documents were found by only a single participating research team. For 7 of the 25 topics, more than half of the known relevant documents were ranked in the top-70 in runs submitted by only a single research team. For another 6 of the 25 topics, between 40 and 50 percent of their relevant documents were ranked in the top-70 by only one team.

These results show a substantial contribution to the relevance pool from each site, with far less overlap than has been typical in previous TREC evaluations. This limited degree of overlap could result from the following factors:

- A preponderance of fairly broad topics for which many relevant documents might be found in the collection. The average of 165 relevant documents per topic is somewhat greater than the value typically seen at TREC (100 or so).
- The limitation of the depth of the relevance judgment pools to 70 documents (100 documents per run have typically been judged in prior TREC evaluations).
- The diversity of techniques tried by the participating teams in this first year of Arabic retrieval experiments at TREC, which could produce richer relevance pools.
- A relatively small number of participating research teams, which could interact with the diversity

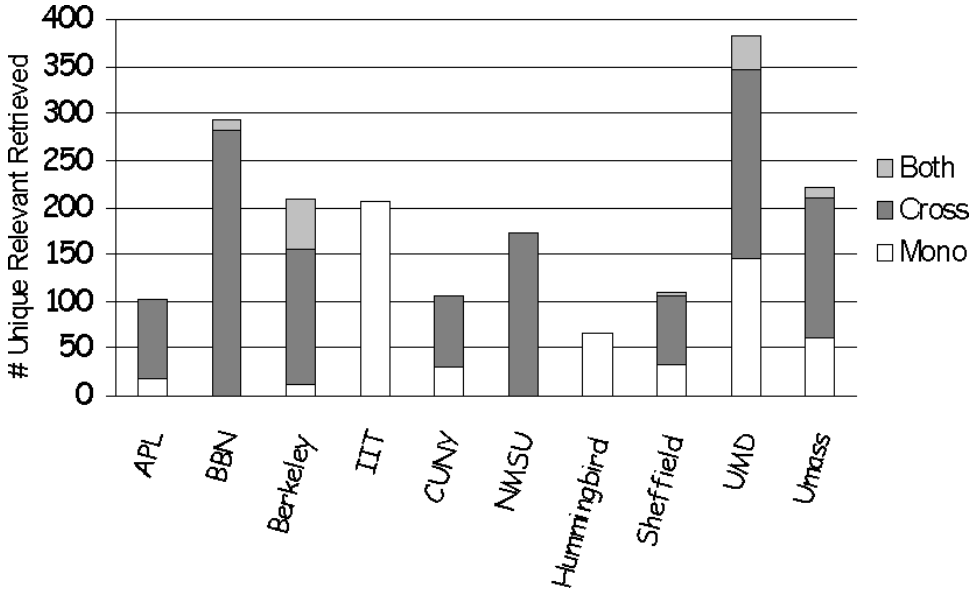


Figure 3: Unique relevant documents, by research team.

of the techniques to make it less likely that another team would have tried a technique that would find a similar set of documents.

The first two factors have occasionally been seen in information retrieval evaluations based on pooled assessment methodologies (TREC, CLEF, and NTCIR) without the high "uniques" effect observed on this collection. We therefore suspect that the dominant factors in this case may be the last two. But until this cause of the high "uniques" effect is determined, relative differences of less than 15% or so in unjudged and post hoc runs using this collection should be regarded as suggestive rather than conclusive. There is, of course, no similar concern for comparisons among judged runs since judgments for their "uniques" are available.

As has been seen in prior evaluations in other languages, manual and monolingual runs provided a disproportionate fraction of the known relevant documents. For example, 33% of the relevant documents that were found by only one team were found only by monolingual runs, while 63% were found only by cross-language runs.

4 Results

Table 1 summarizes the alternative indexing terms, the query languages, and (for cross-language runs) the sources of translation knowledge that were explored by the ten participating teams. All ten participating teams adopted a "bag-of-terms" technique based on indexing statistics about the occurrence of terms in each document. A wide variety of specific techniques were used, including language models, hidden Markov models, vector space models, inference networks, and the PIRCS connectionist network. Four basic types of indexing terms were explored, sometimes separately and sometimes in combination:

Words. Indexing word surface forms found by tokenizing at white space and punctuation requires no language-specific processing (except, perhaps, for stopword removal), but potentially desirable matches between morphological variants of the same word (e.g., plural and singular forms) are

Team	Arabic Terms Indexed				Query Lang	Translation Resources Used			
	Word	Stem	Root	n-gram		MT	Lexicon	Corpus	Translit
BBN		X			A,E	X	X	X	
Hummingbird		X			A				
IIT	X	X	X		A,E	X	X		
JHU-APL	X			X	A,E,F	X			
NMSU	X	X			A,E		X		
Queens	X			X	A,E	X			
UC Berkeley		X			A,E	X	X		
U Maryland	X	X	X	X	A,E	X			X
U Mass	X	X			A,E	X	X		
U Sheffield	X				A,E,F	X			

Table 1: Configurations tested by participating teams.

precluded. As a result, word indexing yielded suboptimal retrieval effectiveness (by the mean uninterpolated average precision measure). Many participating research teams reported results for word-only indexing, making that condition useful as a baseline.

Stems. In contrast to English, where stems are normally obtained from the surface form of words by automatically removing common suffixes, both prefixes and suffixes are normally removed to obtain Arabic stems. Participating teams experimented with stemming software developed at three participating sites (IIT, NMSU, and U Maryland) and from two other sources (Tim Buckwalter and Shereen Khoja).

Roots. Arabic stems can be generated from a relatively small set of root forms by expanding the root using standard patterns, some of which involve introduction of infixes. Stems generated from the same root typically have related meanings, so indexing roots might improve recall (possibly at the expense of precision, though). Although humans are typically able to reliably identify the root form of an Arabic word by exploiting context to choose between alternatives that would be ambiguous in isolation, automatic analysis is a challenging task. Two participating teams reported results based on automatically determined roots.

Character n-grams. As with other languages, overlapping character n-grams offer a useful alternative to techniques based on language-specific stemming or morphological analysis. Three teams explored n-grams, with values of n ranging from 3–6.

Term formation was typically augmented by one or more of the following additional processing steps:

Character deletion. Some Unicode characters, particularly diacritic marks, are optional in Arabic writing. This is typically accommodated by removing the characters when they are present, since their presence in the query but not the document (or vice-versa) might prevent a desired match.

Character normalization. Some Arabic letters have more than one Unicode representation because their written form varies according to morphological and morphotactic rules, and in some cases authors can use two characters interchangeably. These issues are typically accommodated by mapping the alternatives to a single normalized form.

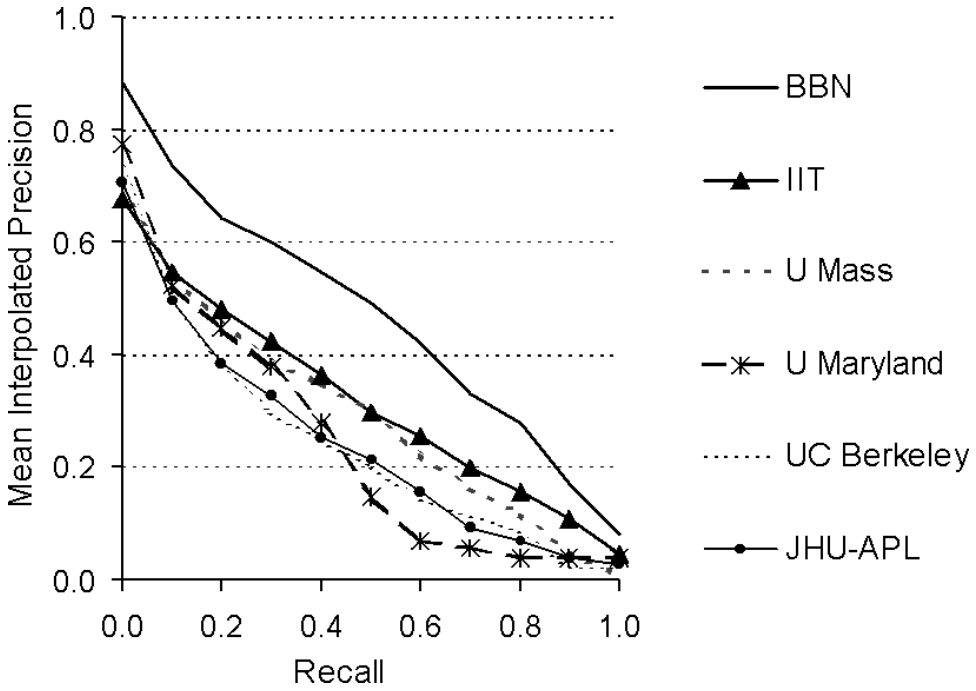


Figure 4: Cross-language retrieval effectiveness, English queries formed from title+description fields, automatic runs.

Stop-term removal. Extremely frequent terms and other terms that system developers judge to be of little use for retrieval are often removed in order to reduce the size of the index. Stop-term removal is most commonly done after stemming or morphological analysis in Arabic because the highly productive morphology would otherwise result in impractically large stopword lists.

Nine of the ten participating research teams submitted cross-language retrieval runs, with all nine using a query-translation architecture. Both of the teams that tried French queries used English as a pivot language for French-to-Arabic query translation, so English-to-Arabic resources were key components in every case. Each team explored some combination of the following four types of translation resources:

Machine Translation Systems. Two machine translation systems were used: (1) a system developed by Sakhr (available at http://tarjim.ajeeb.com, and often referred to simply as "Ajeeb" or "Tarjim"), a system produced by ATA Software Technology Limited (available at http://almisbar.com, and sometimes referred to as "Almisbar" or by the prior name "Al-Mutarjim"). At the time of the experiments, both offered only English-to-Arabic translation. Some teams used a machine translation system to directly perform query translation, others used translations obtained from one or both of these systems as one source of evidence from which a translated query was constructed. A mark in the "MT" column of Table 1 indicates that one or more existing machine translation systems were used in some way, not that they were necessarily used to directly perform query translation.

Translation Lexicons. Three commercial machine readable bilingual dictionaries were used: one marketed by Sakhr (sometimes referred to as "Ajeeb"), one marketed by Ectaco Inc., (typically referred to as "Ectaco"), and one marketed by Dar El Ilm Lilmalayin (typically referred to as "Al Mawrid"). In addition, one team (NMSU) used a locally produced translation lexicon.

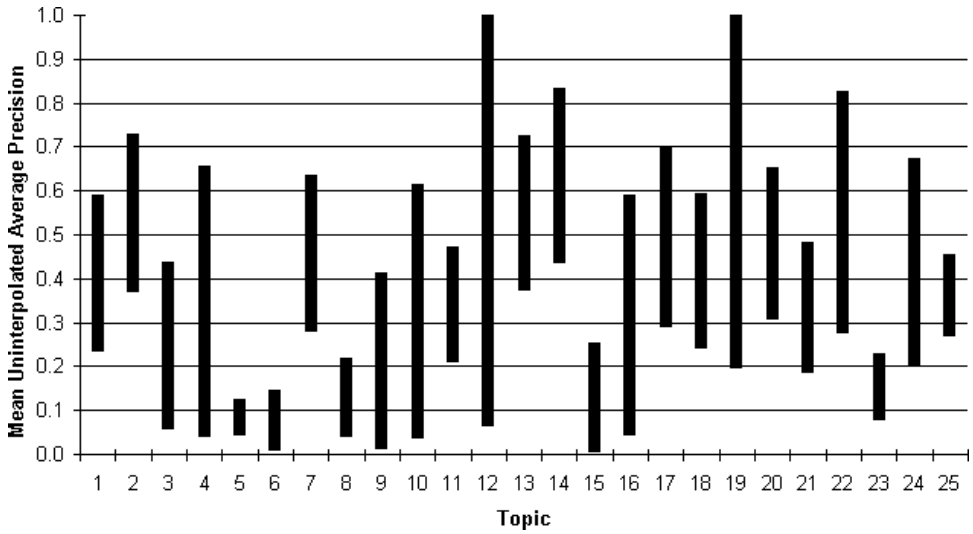


Figure 5: Cross-language topic difficulty, uninterpolated average precision (base of each bar: median over 28 runs, top of each bar: best of the 28 runs).

Parallel Corpora. One team (BBN) obtained a collection of documents from the United Nations that included translation-equivalent document pairs in English and Arabic. Word-level alignments were created using statistical techniques and then used as a basis for determining frequently observed translation pairs.

Transliteration. One team (Maryland) used pronunciation-based transliteration to produce plausible Arabic representations for English terms that could not otherwise be translated.

When multiple alternative translations were known for a term, a number of techniques were used to guide the combination of evidence, including: (1) translation probabilities obtained from parallel corpora, (2) relative term frequency for each alternative in the collection being searched, and (3) structured queries. Pre-translation and/or post-translation query expansion using blind relevance feedback techniques and pretranslation stop-term removal were also explored by several teams.

To facilitate cross-site comparison, teams submitting automatic cross-language runs were asked to submit at least one run in which the query was based solely on the title and description fields of the topic descriptions. Figure 4 shows the best recall-precision curve for this condition by team. All of the top-performing cross-language runs used English queries.

As is common in information retrieval evaluations, substantial variation was observed in retrieval effectiveness on a topic-by-topic basis. Figure 5 illustrates this phenomenon over the full set of cross-language runs (i.e., not limited to title+description queries). For example, half of the runs did poorly on topic 12, which included specialized medical terminology, but at least one run achieved a perfect score on that topic. Topic 5, by contrast, turned out to be problematic for all systems.

No standard condition was required for monolingual runs, so Figure 6 shows the best monolingual run by team regardless of the experiment conditions. Several teams observed surprisingly small differences between monolingual and cross-language retrieval effectiveness. One site (JHU-APL) submitted runs under similar conditions for all three topic languages, and Figure 7(a) shows the resulting recall-precision graphs by topic language. In that case, there is practically no difference between English-topic and Arabic-topic results. There are two possible explanations for this widely observed effect:

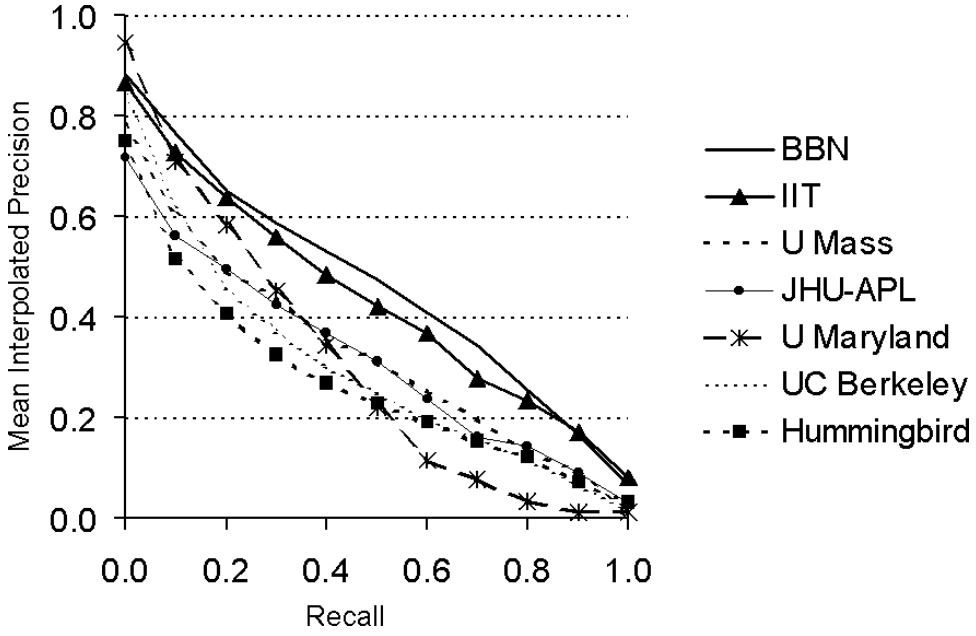


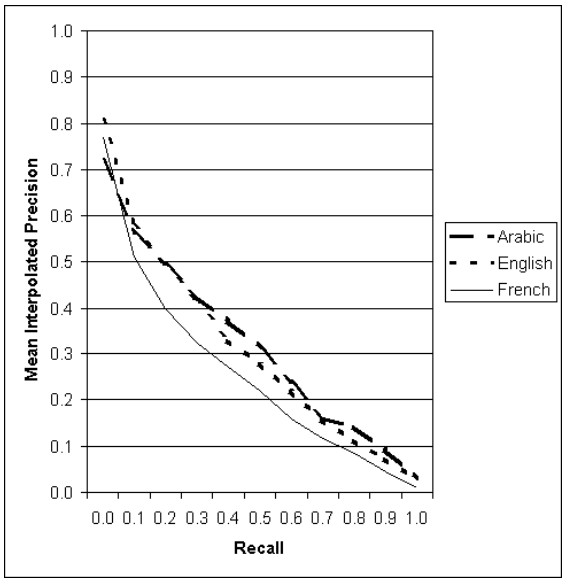
Figure 6: Monolingual retrieval effectiveness, Arabic queries formed from title+description fields (except JHU-APL and UC Berkeley, which also used the narrative field), automatic runs (except U Maryland, which was a manual run).

- No large Arabic information retrieval test collection was widely available before this evaluation, so the monolingual Arabic baseline systems created by participating teams might be improved substantially in subsequent years.
- The 25 topics used in this year's evaluation might represent a biased sample of the potential topic space. For example, relatively few topic descriptions this year included names of persons.

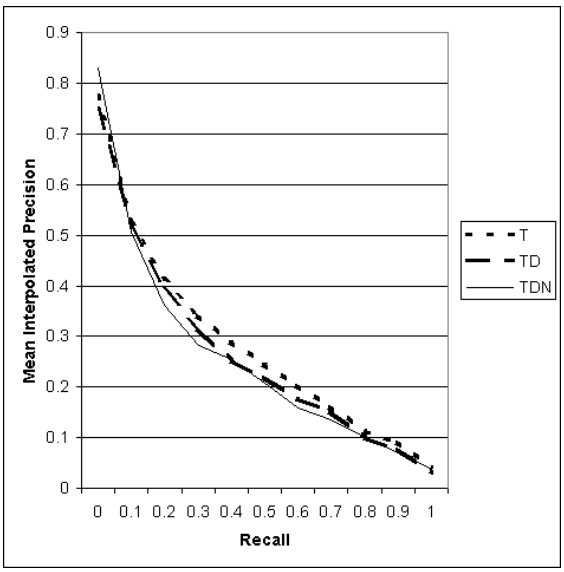
Several teams also observed that longer queries did not yield the improvements in retrieval effectiveness that would normally be expected. One site (Hummingbird) submitted runs under similar conditions for three topic lengths, and Figure 7(b) shows the resulting recall-precision graphs. In this case, longer queries showed no discernible benefit; indeed, it appears that the best results were achieved using the shortest queries! The reasons for this effect are not yet clear, but one possibility is that the way in which the topic descriptions were created may have resulted in a greater concentration of useful search terms in the title field. For example, the title fields contains an average of about 6 words, which is about twice as long as is typical for TREC.

5 Summary and Outlook

The TREC-2001 CLIR track focussed this year on searching Arabic documents using English, French or Arabic queries. In addition to the specific results reported by each research team, the evaluation produced the first large Arabic information retrieval test collection. A wide range of index terms were tried, some useful language-specific processing techniques were demonstrated, and many potentially useful translation resources were identified. In this paper we have provided an overview of that work in a way that will help readers recognize similarities and differences in the approaches taken by the



(a) Topic language effect, title+description+narrative.



(b) Query length effect, Arabic queries.

Figure 7: Comparing runs under comparable conditions (T=title, D=Description, N=Narrative).

participating teams. We have also sought to explore the utility of the test collection itself, providing aggregate information about topic difficulty that individual teams may find useful when interpreting their results, identifying a potential concern regarding the completeness of the pools of documents that were judged for relevance, and illustrating a surprising insensitivity of retrieval effectiveness to query length.

The TREC-2002 CLIR track will continue to focus on searching Arabic. We plan to use 50 new topics (in the same languages) and to ask participating teams to also rerun the 25 topics from this year with their improved systems as a way of further enriching the existing pools of documents that have been judged for relevance. We expect that the result with be a test collection with enduring value for post hoc experiments, and a community of researchers that possess the knowledge and resources needed to address this important challenge.

Acknowledgments

We are grateful to Ellen Voorhees for coordinating this track at NIST and for her extensive assistance with our preliminary analysis, to the participating research teams for their advice and insights along the way, and to Kareem Darwish for his comments on an earlier version of this paper.

References

[1] Carol Peters. *Cross-Language Information Retrieval and Evaluation.* Springer: Lecture Notes in Computer Science: LNCS 2069, Germany, 2001.

[2] Ellen M. Voorhees. Variations in relevance judgments and the mesaurement 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.