CLIR Evaluation at TREC

Donna Harman
National Institute of Standards and Technology
Gaithersburg, Maryland

http://trec.nist.gov

Workshop on Cross-Linguistic Information Retrieval SIGIR 1996

- Paper "Building a Large Multilingual Test Collection from Comparable News Documents" by Páraic Sheridan, Jean Paul Ballerini and Peter Schäuble
- Used Swiss news agency (SDA) data in French, German and Italian

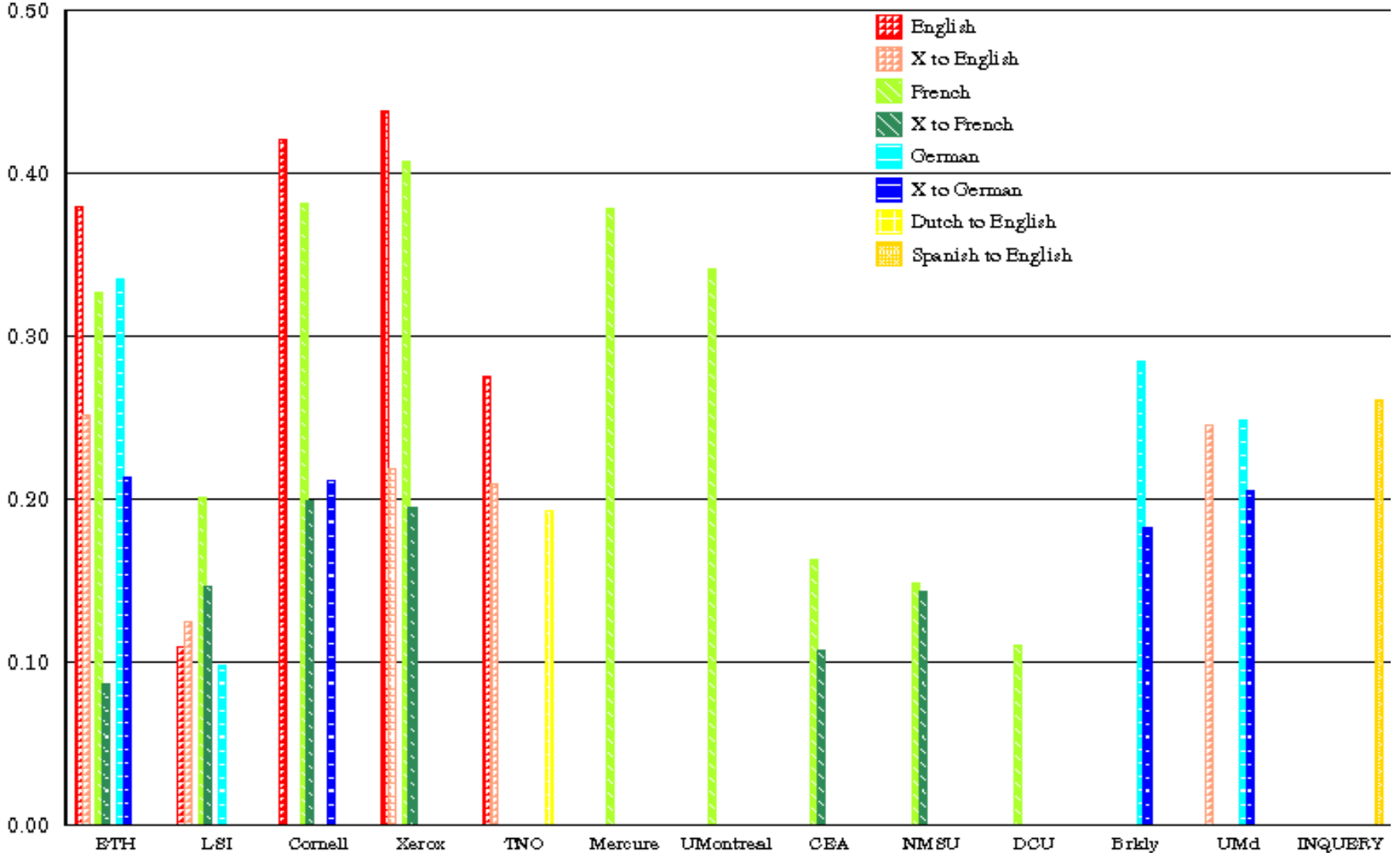
TREC-6 Cross-Language Track

- In cooperation with the Swiss Federal Institute of Technology (ETH)
- Task Summary: retrieval of English, French, and German documents, both in a monolingual and a cross-lingual mode
- Guidelines: ad hoc task guidelines, plus all groups had to submit a monolingual baseline
- Documents:
  - SDA (1988-1990): French (250 MB), German (330 MB)
  - Neue Zürcher Zeitung (1994): German (200 MB)
  - AP (1988-1990): English (759 MB)
- Topics and relevance assessments all done at NIST

TREC-6 Cross-Language Track

- 13 participating groups
  - CEA/DIST/SMTI (France) -E/F
  - Cornell University (USA) - E/F, E/G
  - Dublin City University - F
  - Duke University/U. Colorado/Bellcore - E/F, E/G, F/E, etc.
  - Swiss Federal Institute of Technology (ETH) - E/F, E/G, etc.
  - Xerox Research Centre Europe (France) - E/F
  - IRIT/SIG (France) - F
  - New Mexico State University - E/F
  - TNO/University of Twente - F/E, G/E, Dutch/English
  - University of California, Berkeley - E/G
  - University of Maryland - E/G, G/E
  - University of Massachusetts, Amherst - Spanish/English
  - University of Montreal - F

TREC-6 Cross-Language Results - *revised 01/20/98*



Comments on TREC-6

- Should be viewed as a dry run
- Validates use of "comparable" corpora, but need to extend beyond Swiss newswire
- Problems with creation of multilingual topics at NIST
- Many different approaches used, including simple statistical methods, complex statistical methods, and sophisticated NLP methods
- All methods had problems, but different problems

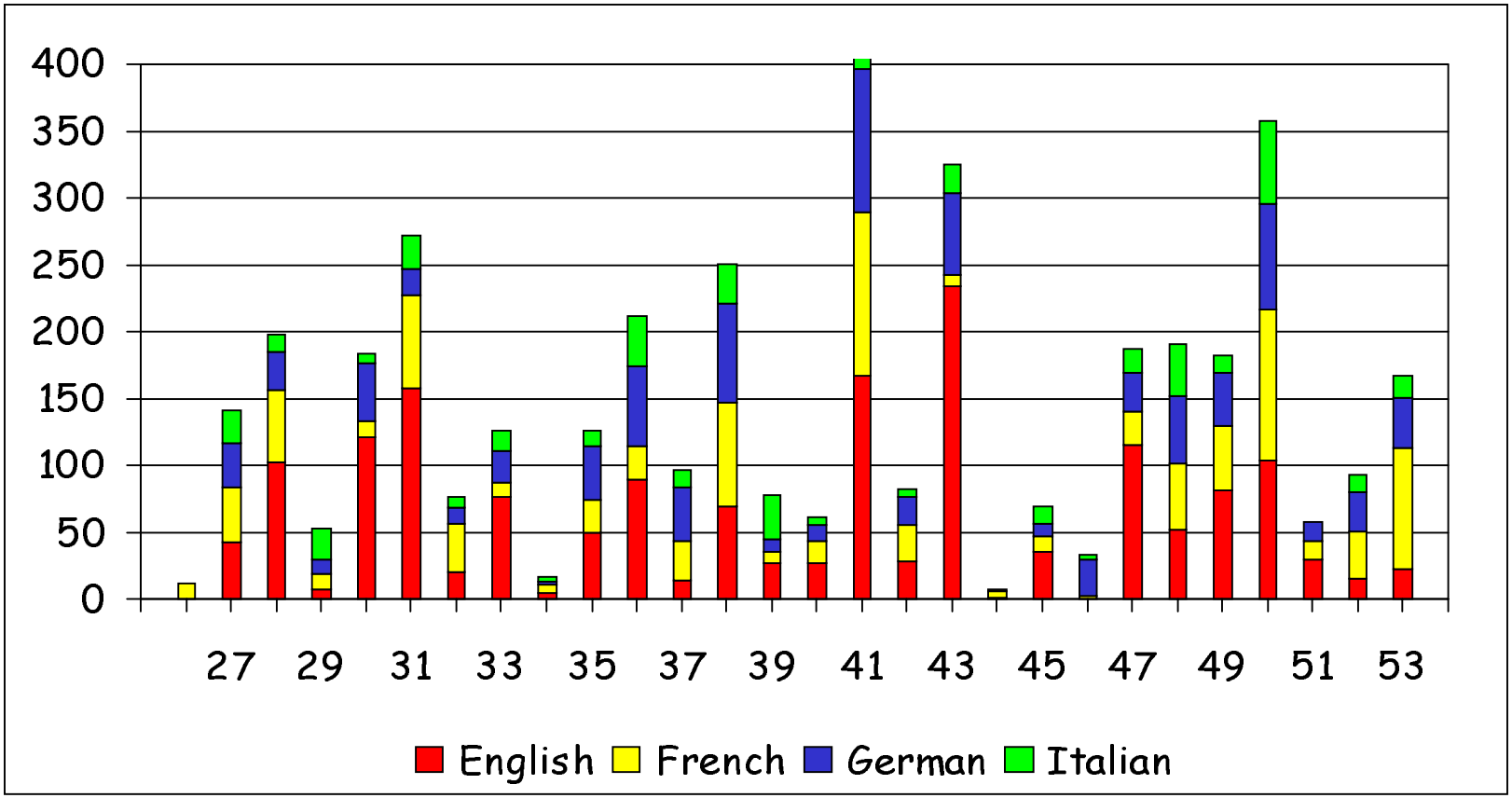
TREC-7 Cross-Language Track

- Task Summary: retrieval of English, French, German and Italian documents; results to be returned as a single multilingual ranked list
  - Addition of Italian SDA (1989-1990), 90 MB
  - Addition of a subtask of 31,000 structured German social science documents (GIRT)
- Topics and relevance judgments done at 4 sites
  - English: NIST
  - French: EPFL Lausanne
  - German: Informationszentrum Sozialwissenschaften
  - Italian: CNR Pisa

TREC-7 Cross-Language Track

- 9 participating groups
  - CEA – E/F
  - Eurospider Information Technology AG - F/E/G/I
  - IBM T.J. Watson Research Center – F/E/G/I
  - TextWise Inc. – F/E
  - Los Alamos National Laboratory (USA) – E/G
  - TNO/University of Twente – F/E/G/I
  - University of California, Berkeley – F/E/G/I
  - University of Maryland – F/E/G/I
  - University of Montreal – F/E

TREC-7 Number Relevant by Language



Comments on TREC-7

- Much improved topic creation and relevance judgments from TREC-6
- Confounded two different stages of search
  - cross-language retrieval
  - merging of results from multiple languages
- Brought to light some major discussion of what the translation of a topic should be
- Relevance assessments done across 4 sites

French Translation

<num> Number: CH13

<F-title> Dette de la Confédération

<F-desc> Description:
Comment la dette de la Confédération est-elle couverte?

<E-title> The Confederation's public debt

<E-desc> Description:
How is the Confederation's public debt covered?

<num> Number: 46

<E-title> The Swiss Confederation's public debt

<E-desc> Description:
How is the Swiss Confederation's public debt being paid?

<E-narr> Narrative:
Switzerland is among the least indebted countries in the world. The tax on assets is the Confederation's main source of income. However the major part of the public debt is covered by the equivalent of U.S. Treasury bonds. Relevant documents should be found that describe this method of covering the debt or other methods that are being used.

German Translation

GT07

Title: Umschuldungsabkommen für Polen

Description: Welche Berichte und Analysen gibt es zu
der Umschuldung und den entsprechenden Abkommen für Polen?

English title: Conversion of debt for Poland

English description: Which reports and studies are given
concerning the conversion of debt for Poland?

<num> Number: 38
<E-title> Conversion of debt for Poland

<E-desc> Description:
What reports and studies are there concerning the conversion of debt
for Poland?

<E-narr> Narrative:
Relevant documents deal with the negotiations and their results
concerning the conversion of debt for Poland. Reports on the indebtedness
of Poland and the activities aiming on the conversion of the debt within
the context of bilateral and international agreements are also relevant.

TREC-8 Cross-Language Track

- Tasks, documents and topic creation similar to TREC-7
- 12 participating groups
  - Claritech (USA) - F/E/G/I
  - Eurospider Information Technology AG - F/E/G/I, GIRT
  - IBM T.J. Watson Research Center - F/E/G/I
  - IRIT/SIG (France) - F/E/G/I
  - Johns Hopkins University APL (USA) - F/E/G/I
  - MNIS-Textwise Labs - F/E
  - New Mexico State University - F/E/G/I
  - Sharp Laboratories of Europe Ltd (UK) - I/E
  - TNO/University of Twente - F/E/G/I
  - University of California, Berkeley - F/E/G/I, GIRT
  - University of Maryland - F/E/G/I
  - University of Montreal - F/E

Comments on Topic Translation in TREC-8

- In TREC-7 we tried to resolve these issues at NIST; in TREC-8 there was more time for cooperation. Additionally there was a final pass made on the topic translations by the University of Hildesheim (thanks to Prof. Christa Womser-Hacker).
- Note that the degree of literalness in the translations of topics has a major effect on the results of cross-language retrieval
- Topics should clearly not be literal translations, but should conform to the idea that the translation would reflect what a user of a different language (and of a different culture) would input as a question

CLIR in TREC-9

- Documents
  - Hong Kong Commercial Daily, Hong Kong Daily News, Takungpao: all from 1999 and about 260 MB total
- 25 new topics built in English; translations made to Chinese
- All topics and relevance judgments done at NIST

TREC-9 CLIR Participants

- BBN Technologies (USA)
- The Chinese University of Hong Kong
- Fudan University (P. R. China)
- IBM T.J. Watson Research Center (USA)
- Johns Hopkins University (USA)
- KAIST (Korea)
- Microsoft Research, Beijing Lab
- MNIS-TextWise Labs (USA)
- National Taiwan University
- Queens College, CUNY (USA)
- RMIT University (Australia)
- Telcordia Technologies, Inc. (USA)
- Trans-EZ Inc. (USA)
- University of California, Berkeley
- University of Maryland
- University of Massachusetts, Amherst