

# Report from TREC-9

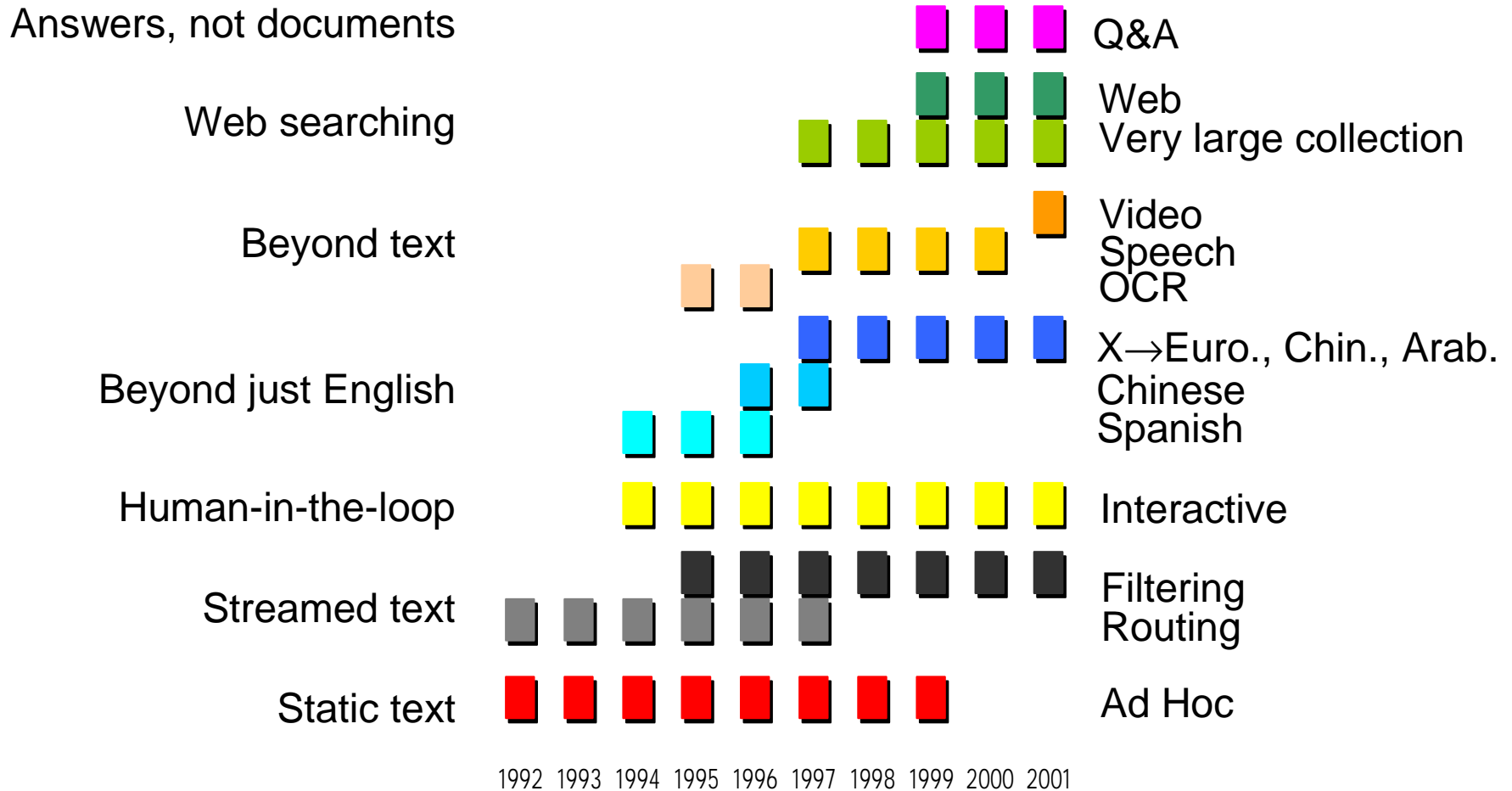
Donna Harman, Ellen Voorhees

Retrieval Group

Information Access Division

National Institute of Standards and  
Technology

# TREC Tasks



# TREC-9 Tracks

- Cross-language (English to Chinese)
- Filtering
- Interactive
- Query
- Question Answering
- Spoken Document Retrieval
- Web

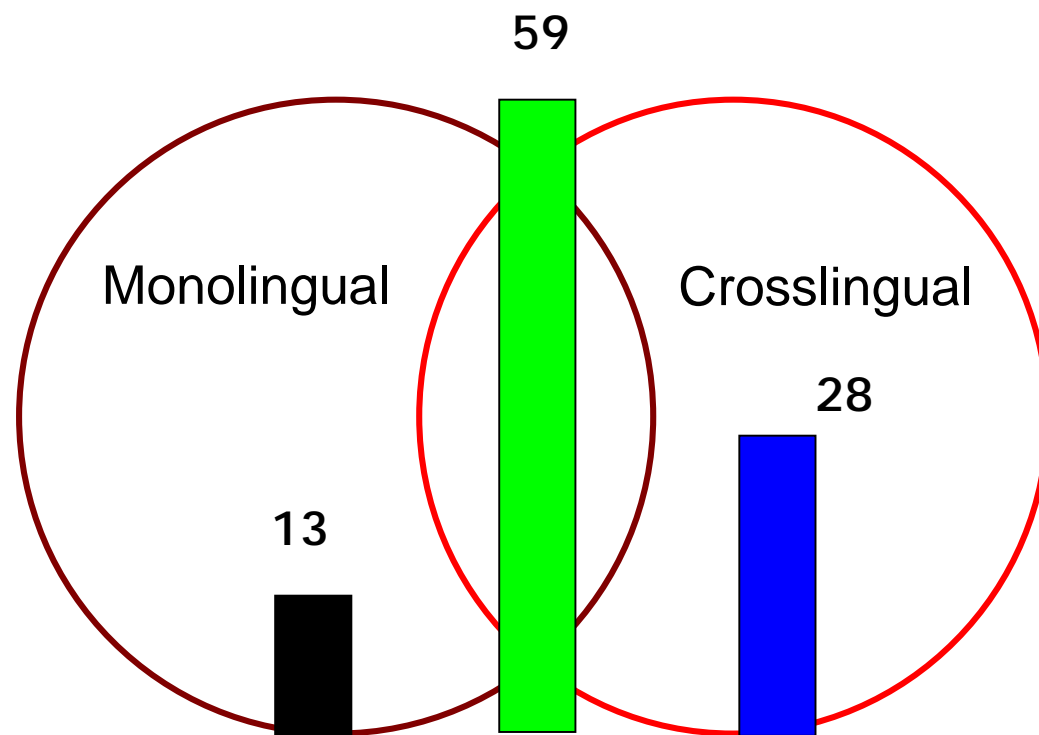
# Cross Language Track

- Task: ad hoc search for documents written in one language using topics in another language
  - 25 topics in English created by bilingual assessors; Chinese version also available
  - 126,937 documents; 188 MB in BI G5
  - Hong Kong newspapers donated by Wiser Ltd.
    - Hong Kong Commercial Data (Aug 98-Jul 99)
    - Hong Kong Daily News (Feb 99-July 99)
    - Takongnao (Oct 98-Mar 99)

# Relevance Judgments

- Judged highest priority mono- and cross-lingual run from each group
  - 39 cross (75%) / 13 mono (25%)
  - 51 auto / 1 manual (Thank you, Berkeley!)
- Added top 50 documents from each judged run to the pool
- Mean actual pool size = 598 (39% of max) within expected range

# % Contributions to Pool by Run Type (Relevant documents)



# Participants

BBN Technologies

Fudan University

IBM T.J. Watson Research Center

Johns Hopkins University

Korea Advanced Institute of Science and  
Technology

Microsoft Research, China

MNI S-TextWise Labs

National Taiwan University

# More participants

Queens College, CUNY

RMIT University

Telcordia Technologies, Inc.

The Chinese University of Hong Kong

Trans-EZ Inc.

University of California at Berkeley

University of Maryland

University of Massachusetts



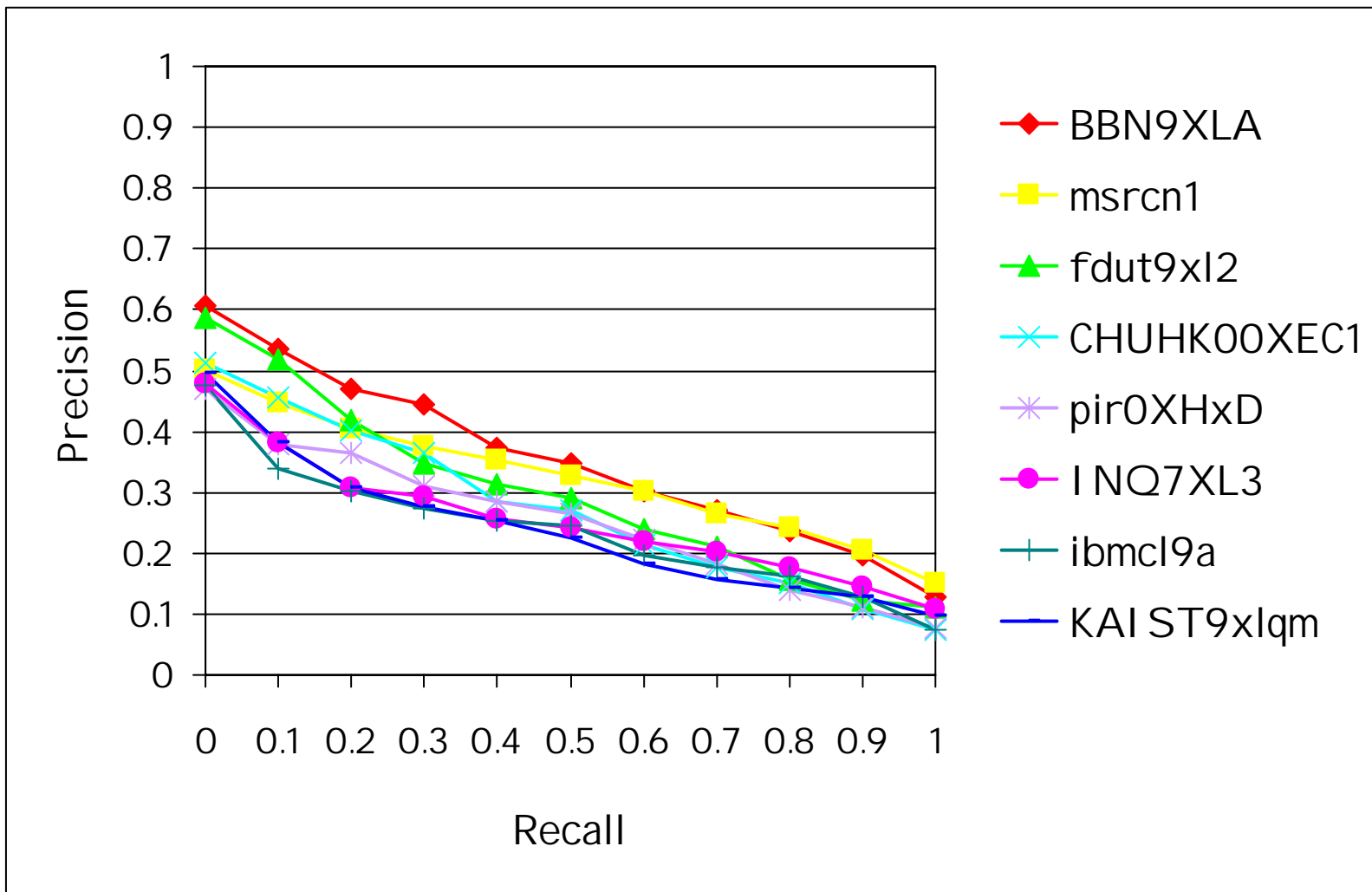
# Resources: dictionaries/word lists

- LDC English - Mandarin word list (~120,000 pairs)
- Chinese-English Translation Assistance (CETA) dictionary
- KingSoft online bilingual dictionary
- WordNet
- other local (proprietary) dictionaries

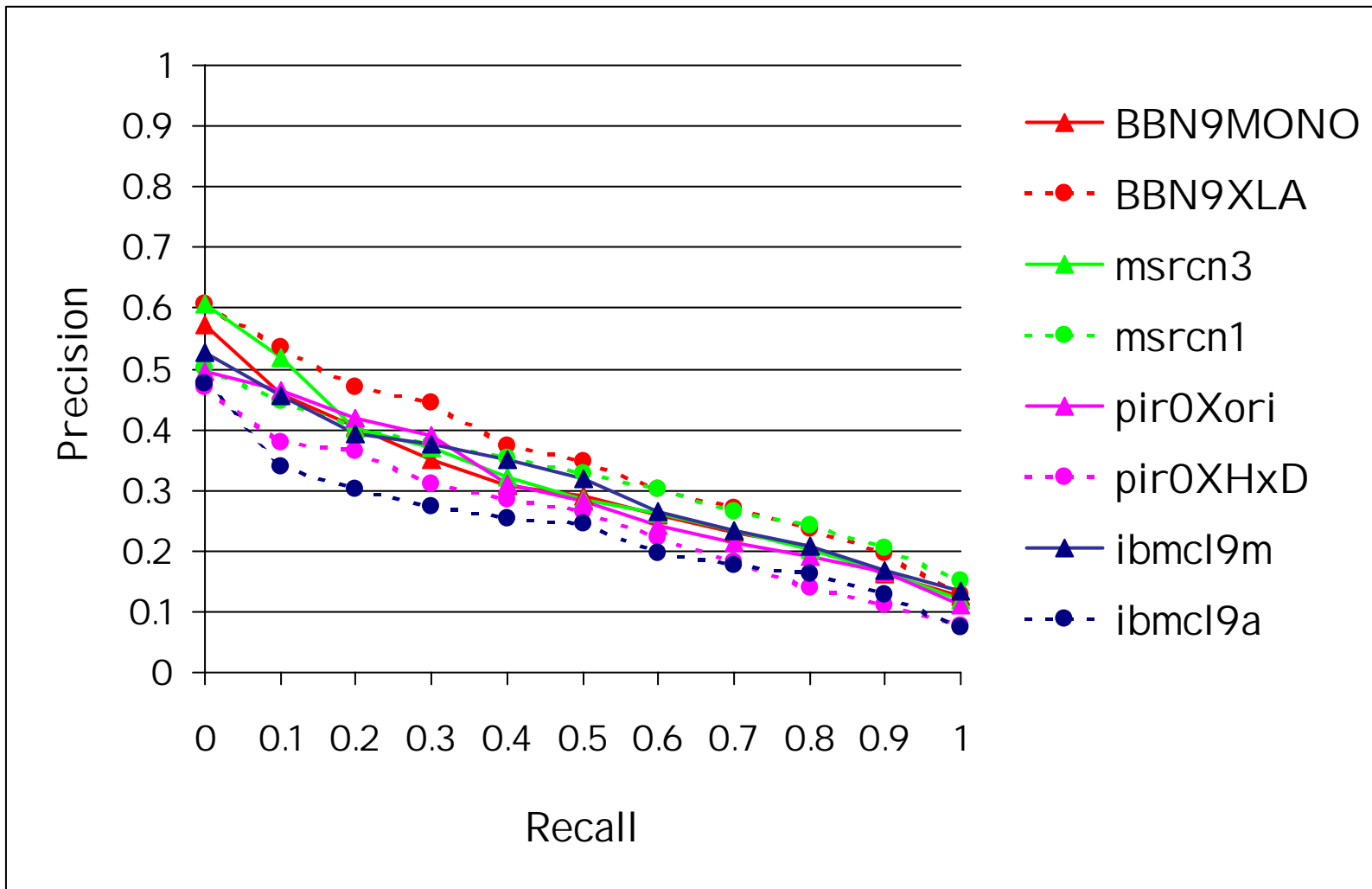
# Resources: software & services

- MT
  - HuaJian MT system
  - IBM AlphaWorks translation server
  - Alis Gist-in-Time MT system
- English analysis
  - InXight LinguistX (English linguistic analysis)
  - Apple Pie parser
  - Brill's POS tagger
- Chinese analysis/conversion
  - Various Chinese segmenters (e.g., NMSU's ch\_seg)
  - BI G5->GB converters (e.g., NJStar's)
- Miscellaneous
  - CMU's WEAVER translation-pair extraction
  - Yahoo search

# English to Chinese Results

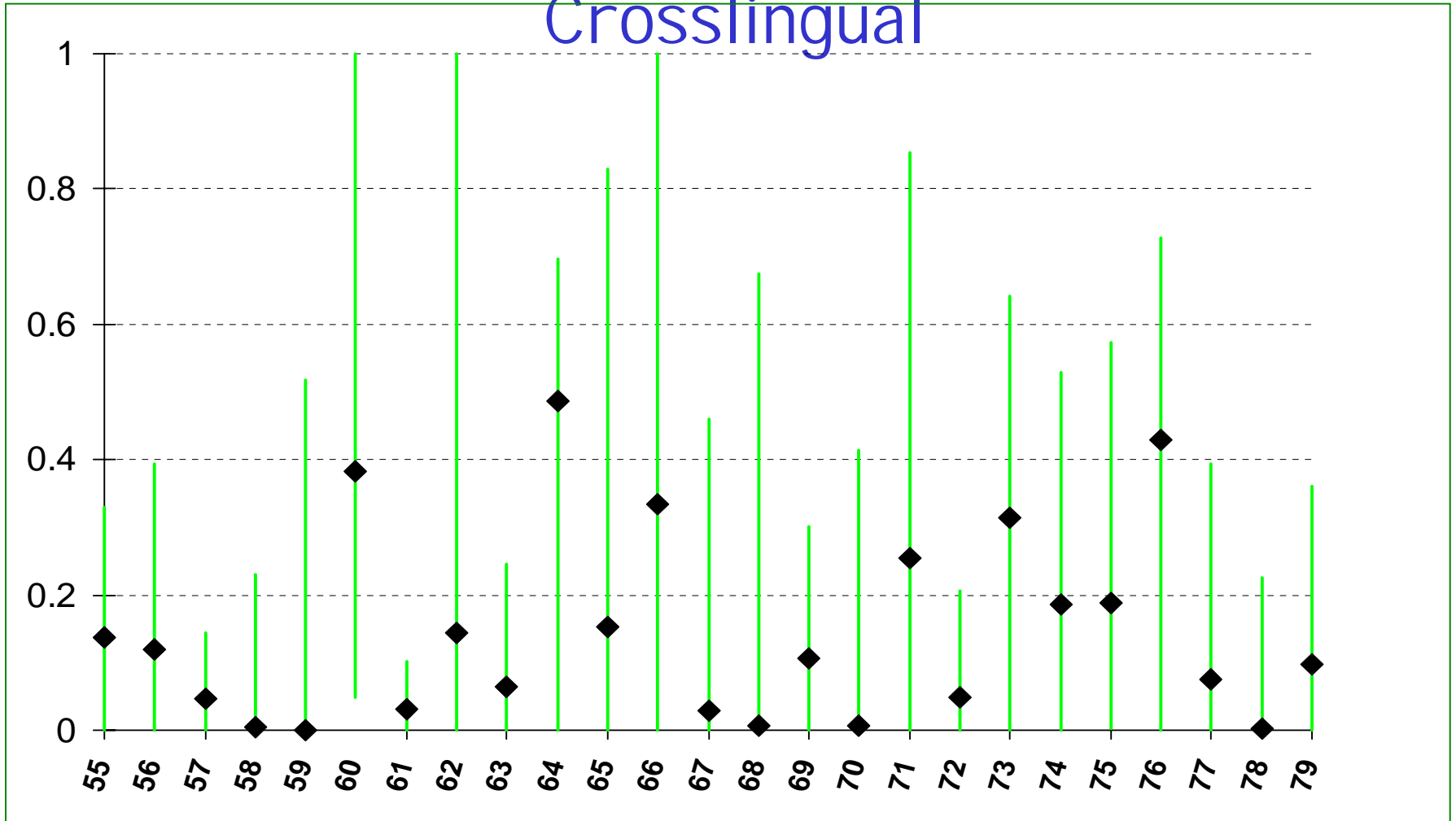


# Cross-language vs. Monolingual



# Average Precision by Topic:

## Crosslingual



# What was learned from the Chinese CLIR track?

- Many approaches to English to Chinese topic translation, including use of various dictionaries, word lists, parallel text, and commercial MT systems
- Extensive set of Chinese retrieval experiments performed ranging from various n-gram methods to word based to complete language modeling
- Because of the tight focus of this track, cross-system comparison is possible

# TREC 2001

- Cross language
  - Chinese → NTCIR workshop (NII, Japan)
  - TREC task will be English, French→Arabic
- Filtering track using new Reuters corpus
- Interactive to investigate live web
- Expanded web and QA tracks
- New video track

# trec.nist.gov

## Text REtrieval Conference (TREC)

*...to encourage research in information retrieval  
from large text collections.*

Overview

Frequently Asked  
Questions

Publications

Data

Information  
for Active  
Participants



Contact  
Information

Past TREC  
Results