CLEF 2004 Cross-Language Spoken Document Retrieval Track

Marcello Federico[1], Gareth Jones[2]

[1] **ITC-irst, Italy**

[2] **Dublin City University, Ireland**

Overview

- The Cross Language Spoken Document Retrieval (CL-SDR) track aims to evaluate CLIR systems on noisy automatic transcripts of spoken documents.
- The CLEF 2004 CL-SDR track relies once again on data prepared by NIST for the TREC 8-9 SDR tracks.
- The original English short queries were manually translated to other, e.g. French or German, for CL-SDR.
- Retrieval is performed on automatic transcriptions made available by NIST, and generated by different speech recognition systems.

Overview

- For the CLEF 2004 CL-SDR task an unknown story boundary condition.
- For the CLEF 2003 CL-SDR track the transcription was manually divided into individual story units. This year participants were provided with unsegmented transcripts.
- For each query, systems had to produce a ranked list of relevant stories, based on identifying a complete news show and a time index within the news show.
- Relevance is assessed by checking if the provided time indexes fall inside the manually judged relevant stories.

Data Specification

- he document collection consists of 557 hours of American-English news recordings broadcast by: ABC, CNN, Public Radio International (PRI), and Voice of America (VOA) between February and June 1998.
- Spoken documents are accessible through automatic transcriptions produced by NIST and other sites, which participated in the TREC-9 SDR track.
- Transcripts are provided with and without story boundaries, for a total of 21,754 stories.

Data Specification

- For the application of blind relevance feedback, participants were allowed to use parallel document collections, such as those available through the Linguistic Data Consortium.

Data Specification

- The CLEF topics are based on 100 English topics generated for TREC-8 and TREC-9 SDR.
- The original TREC relevance assessments were used.
- For CLEF the topics were translated by native speakers into Dutch, Italian, French, German, and Spanish.
- Scoring software was made available both for the known and unknown story boundary conditions.

Data Specification

The following evaluation conditions were specified:

- Primary Conditions (mandatory for all participants):
  – Monolingual IR on NIST transcripts, no parallel data.
  – Bilingual IR from French/German on NIST transcripts, no parallel data.
- Secondary Conditions (optional):
  – Bilingual IR from French/German, on NIST transcripts, with parallel data.
  – Bilingual IR from any language, all transcripts,with parallel data.

Participants

- University of Chicago
- ITC-irst