# Preliminary Work on Cross-Language Spoken Document Retrieval

M. Federico [1], N. Bertoldi [1] and G. Jones [2]

[1] ITC-irst - Centro per la Ricerca Scientifica e Tecnologica, Italy

[2] University of Exeter, United Kingdom

# Objectives

- Evaluation of CLIR systems on noisy automatic transcripts of spoken docs
- Low-cost development of a benchmark using available TREC SDR collections
- Decide upon future directions: benchmark or evaluation?

# CL-SDR Benchmark

- **Target collection:**
  - Automatic transcripts of 389 hours of broadcast news (Feb.-Jun. '98)
  - 21,754 manually segmented stories - Known Story Boundaries
  - NIST/BBN baseline ASR: 26,7% WER (B2K TREC'99, B1K TREC'00)
- **Test data:**
  - Original topics: 49 of TREC'99 + 50 of TREC'00
  - Translations by ITC-irst: Italian, Dutch, German, French and Spanish
  - Description (short)
  - Relevant documents: 1818 + 2216
- **Parallel Corpus:**
  - 314,697 texts North American News (Sep. '97 - Apr. '98):

# CLIR System Italian-English

- **Text Preprocessing:**
  - baseform on Italian and stemming on English
  - Stop-term removal
  - Proper names and number recognition
- **CLIR Models:**
  - Q-D Model (retrieval): Okapi + Language Model
  - Q-T Model (translation): $N$-best translation decoder
- **Blind Relevance Feedback:**
  - on target collection (t_BRF)
  - on parallel collection (p_BRF)
  - first on parallel collection, then on target collection (p+t_BRF)

# Monolingual SDR Experiments

| type | year | Trs | WER(%) | mAvPr |
|---|---|---|---|---|
| IRST | '99 | B2K'99 | 26.7 | .5640 |
| CUHTK | '99 | B2K'99 | 26.7 | .5302 |
| Sheffield | '99 | B2K'99 | 26.7 | .5335 |
| LIMSI | '99 | B2K'99 | 26.7 | .4839 |
| IRST | '00 | B1K'00 | 26.7 | .4372 |
| CUHTK | '00 | B1K'00 | 26.7 | .4831 |
| Sheffield | '00 | B1K'00 | 26.7 | .4620 |

Table 1: mAvPr results on TREC'99 & '00 SDR task.

Participants in TREC'00 SDR task developed their system on TREC'99 data.

# Cross-Language SDR Experiments

| type | year | $N$-best | base | t_BRF | p_BRF | p+t_BRF |
|---|---|---|---|---|---|---|
| IRST | '99 | 1 | .2621 | .3345 | .3638 | .3772 |
| IRST | '99 | 5 | .2618 | .3347 | .3571 | .3922 |
| IRST | '99 | 10 | .2626 | .3338 | .3562 | .3898 |
| SYSTRAN | '99 | 1 | .2366 | .3094 | .3803 | .3946 |
| IRST | '00 | 1 | .2018 | .2694 | .2597 | .2892 |
| IRST | '00 | 5 | .1989 | .2622 | .2673 | .2750 |
| IRST | '00 | 10 | .1989 | .2622 | .2673 | .2750 |
| SYSTRAN | '00 | 1 | .2255 | .2786 | .2943 | .3139 |

Table 2: `mAvPr` results on TREC'99 & '00 CLSDR task.

# Current view

- Setting-up a benchmark for CL-SDR is feasible if one can exploit an existing evaluation framework (TREC SDR)
  - Major work consists of translating topics into several languages
- Organizing a CL-SDR evaluation is very expensive, for sure not affordable by our organisations
- We will probably continue to develop benchmarks to be made publically available
- We plan to further develop our research using this benchmark

# Discussion about CLEF

- There is room for improvement in the here presented task:
  - major gap between monolingual vs. cross-language SDR
  - minor gap between manual vs. automatic transcripts IR
- Increasing difficulty of the task
  - work with unknown story-boudary condition? (cheap to organize)
  - move to the TDT collection and work with a similar task? (less cheap “)
  - work on topic detection with the TDT collection? (less cheap “)