The CLEF 2003 Cross-Language Spoken Document Retrieval Track

Marcello Federico* Gareth Jones†

Abstract

The current expansion in collections of natural language based digital documents in various media and languages is creating challenging opportunities for automatically accessing the information contained in these documents. This paper describes the CLEF 2003 track investigation of Cross-Language Spoken Document Retrieval (CLSDR) combining information retrieval, cross-language translation and speech recognition. The experimental investigation is based on the TREC-8 and TREC-9 SDR evaluation tasks, augmented to form a CLSDR task. The original task of retrieving English language spoken documents using English request topics is compared with cross-language retrieval using French, German, Italian, Spanish and Dutch topic translations.

1 Introduction

In recent years much independent research has been carried out on multimedia and multilingual retrieval. The most extensive work in multimedia retrieval has concentrated on spoken document retrieval from monolingual (almost exclusively English language) collections, generally using text search requests to retrieve spoken documents. Speech recognition technologies have made impressive advances in recent years and these have proven to be effective for indexing spoken documents for spoken document retrieval (SDR). The TREC SDR track ran for 4 years from TREC-6 to TREC-9 and demonstrated very good performance levels for SDR [2]. In parallel with this, there has been much progress in cross-language information retrieval (CLIR) as exemplified by the CLEF workshops. Good progress in these separate areas means that it is now timely to explore integrating these technologies to provide multilingual multimedia IR systems.

Following on from a preliminary investigation carried out as part of the CLEF 2002 campaign, a Cross-Language Spoken Document Retrieval track was organized for CLEF 2003. Developing a completely new task for this track was beyond available resources, and so the track built on the work from the CLEF 2002 pilot track [1] and is mainly based on existing resources. The existing resources, kindly made available by NIST, were used at for the TREC 8 and 9 monolingual SDR tracks [2]. Hence, the track results are closer to a benchmark than to a real evaluation.

In particular the NIST collection consists of:

- a collection of automatic transcripts (557 hours) of American-English news recordings broadcasted by ABC, CNN, PRI (Public Radio International), and (VOA) Voice of America made between February and June 1998. Transcripts are provided both with unknown story boundaries, and with known story boundaries (21,754 stories).
- two sets of 50 English topics (one each from TREC 7 and TREC 8) either in terse or short format.
- manual relevance assessments

*ITC-irst - Centro per la Ricerca Scientifica e Tecnologica, I-38050 Povo, Trento, Italy.
†Department of Computer Science, University of Exeter, U.K.

- scoring software for the known/unknown story boundary condition

The TREC collections have been extended to a CLSDR task by manuall translating with the short topics into five European languages: Dutch, Italian, French, German, and Spanish.

Track Specifications

The track aimed at evaluating CLIR systems on noisy automatic transcripts of spoken documents with known story boundaries. The following specifications were defined about data and resources participants were allowed to use for development and evaluation purposes.

Development data (from TREC 8 SDR)

a Document collection: the B1SK Baseline Transcripts collection with known story boundaries made available by NIST.

b Topics: Short topics in English, French, German, Italian, Spanish and Dutch made available by ITC-irst.

c Relevance assessments: Topics-074-123.

d Parallel document collections (optional): available through LDC.

Evaluation data (from TREC 9 SDR)

a Document collection: the B1SK Baseline Transcripts collection with known boundaries made available by NIST.

b Topics: Short topics in English, French, German, Italian, Spanish and Dutch.

c Relevance assessments: Topics-124-173

d Parallel document collections (optional): available through LDC.

Primary Conditions (mandatory for all participants)

- Monolingual IR without using any parallel collection (contrastive condition).
- Bilingual IR from French or German.

Secondary Condition (optional)

- Monolingual IR using any available parallel collections.
- Bilingual IR from other languages.

2 Participants

Four research groups registered to participate in this track:

- University of Alicante (Spain)
- Johns Hopkins University (USA)
- University of Exeter (U.K.)
- ITC-irst (Italy)

Official run	Site	Query	mAvPr
resultsEnconexp	UAlicante	EN	.3563
resultsEnsinexp	UAlicante	EN	.2943
aplspenena	JHU/APL	EN	.3184
exeengpl1.5	UExeter	EN	.3824
exeengpl3.5	UExeter	EN	.3696
Mono-brf	ITC-irst	EN	.3944
resultsFRconexp	UAlicante	FR	.2846
resultsFRsinexp	UAlicante	FR	.1648
aplspfrena	JHU/APL	FR	.1904
exefrprnsys1.5	UExeter	FR	.2825
exefrprnsys3.5	UExeter	FR	.2760
fr-en-1bst-brf-bfr	ITC-irst	FR	.2281
fr-en-sys-brf-bfr	ITC-irst	FR	.3064
aplspdeena	JHU/APL	DE	.2206
exedeprnsys1.5	UExeter	DE	.2744
exedeprnsys3.5	UExeter	DE	.2681
de-en-dec-1bst-brf-bfr	ITC-irst	DE	.2676
de-en-sys-brf-bfr	ITC-irst	DE	.2880
aplspitena	JHU/APL	IT	.2046
exeitprnpro1.5	UExeter	IT	.3011
exeitprnsys1.5	UExeter	IT	.2998
it-en-1bst-brf-bfr	ITC-irst	IT	.2347
it-en-sys-brf-bfr	ITC-irst	IT	.3218
aplspesena	JHU/APL	ES	.2395
exespprnpro1.5	UExeter	ES	.3151
exespprnsys3.5	UExeter	ES	.3077
es-en-1bst-brf-bfr	ITC-irst	ES	.2746
es-en-sys-brf-bfr	ITC-irst	ES	.3555
aplspnlena	JHU/APL	NL	.2269

Table 1: mAvPr results of CLSDR track at CLEF 2003

3 Results and Discussion

Table 1 shows a summary of average precision results for the participants official submissions. It is clearly not possible to analyze the effectiveness of the methods employed by the participants ahead of the workshop. However, it is clear that some methods are on average proving more effective than others, even between separate runs submitted by individual groups. We expect that the methods underlying successful and unsuccessful results will be described by the participants in their individual papers.

We look forward to discussing the approaches taken by the participants at the workshop. It is hoped that these will suggest some definite directions for further research in CLIR for noisy document data.

References

[1] G. J. F. Jones and M. Federico. CLEF 2002 Cross-Language Spoken Document Retrieval Pilot Track Report. In *Proceedings of the CLEF 2002: Workshop on Cross-Language Information Retrieval and Evaluation*, Rome, September 2002. Springer Verlag.

[2] J. S. Garafolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access*, pages 1–20, Paris, 2000.