The CLEF 2003 Cross-Language Spoken Document Retrieval Track

Marcello Federico[1] and Gareth J. F. Jones[2]

[1] ITC-irst, Trento, Italy
[2] Dept. of Computer Science, University of Exeter, U.K.

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 information 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 information retrieval. The most extensive work in multimedia information 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 (CLSDR) 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 [3] and is mainly based on existing resources. These existing resources, kindly made available by NIST, were used for the TREC 8 and 9 monolingual SDR tracks [2]. Hence, the track results are closer to a benchmark than to a real evaluation.

C. Peters et al. (Eds.): CLEF 2003, LNCS 3237, pp. 646–652, 2004.

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.
- scoring software for the known/unknown story boundary condition.

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

2 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 the data and resources that 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: 50 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: 50 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.

Table 1. `mAvPr` results of CLSDR track at CLEF 2003

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

3 Participants

Four research groups participated in this track:

University of Alicante (Spain) in addition to the mandatory monolingual run, this site submitted two runs with French as source language [5] . The system used performs query translation by means of several commercial off-the-shelf machine translation (MT) systems and performs query-document matching at the level of passage rather than of full document. These submissions adapted their existing document splitting algorithm developed for text data containing punctuation in order to identify passages in spoken data without punctuation marks on the basis of pauses contained in the transcripts. Finally, query expansion was just performed on the target collection.

Johns Hopkins University (USA) submitted one run for all available source languages: Dutch, French, German, Italian, and Spanish [6]. Their system employed n-gram decomposition for collection indexing, query translation, and query-document matching. Document retrieval was performed with a statistical language model. In particular, 5-grams were used in all the official bilingual runs. Query expansion just exploited the target collection.

University of Exeter (UK) submitted two runs for the following source languages: French, German, Italian, and Spanish [4]. The system used applied commercial MT systems for query translation and employed an Okapi retrieval method exploiting standard text preprocessing. In particular, query expansion is performed by using a parallel collection which is not truly contemporary to the searched documents.

ITC-irst (Italy) submitted two runs for the following languages: French, German, Italian, and Spanish [1]. The system used featured a statistical retrieval model integrating retrieval scores over multiple query translations. Query-document retrieval scores are computed with two methods: a statistical language model and an Okapi derived formula. Finally, the ITC-irst system employed a parallel collection for query expansion.

4 Results and Discussion

An overview of all submitted runs is reported in Table 1, which also shows performance in terms of average precision. Precision/recall plots of the primary condition runs are shown in Figures 1-3.

Interestingly, the ranking resulting from the contrastive monolingual run is almost preserved in both primary bilingual runs. In particular, the monolingual run shows performance by the pure retrieval systems, disregarding both the translation component and the query expansion on parallel corpora. However, it must be noticed that the latter feature was only exploited by the systems of U. Exeter and ITC-irst.

An interesting comparison between the system performance across the two conditions is given by the plot in Figure 4, which shows the ratio of mean-average-precisions between the bilingual and monolingual runs for different source languages. This plot points out that systems having better monolingual performance also show, in general, better bilingual retrieval performance. An exception was U. Alicante whose French-English shows the best ratio, despite the fact that its monolingual performance (`mavpr` .3563) was significantly lower than the best one (`mavpr` .3964 by ITC-irst). Comparing the methods used to achieve these results by U. Alicante and U. Exeter, it can be noted that Exeter use only a single MT system (Systran) to obtain this result, whereas Alicante use a combination of three MT systems (Babelfish (a version of Systran)), Power Translator and Free-translator. It may be that Alicante's better relative performance is achieved due to greater coverage and possibly better selection of the translated terms arising from the use of multiple resources.

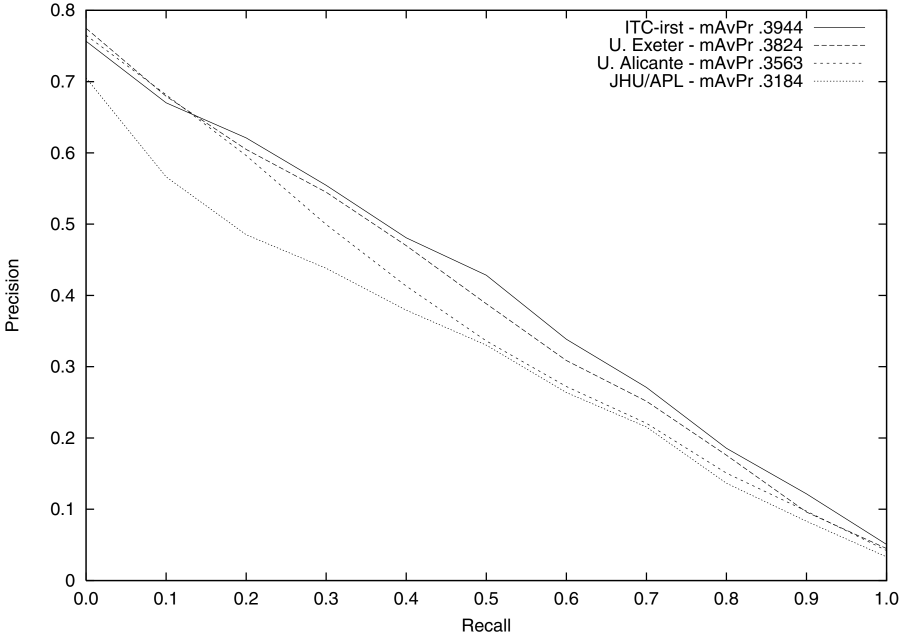

Fig. 1. Precision vs. recall of monolingual runs (primary condition)

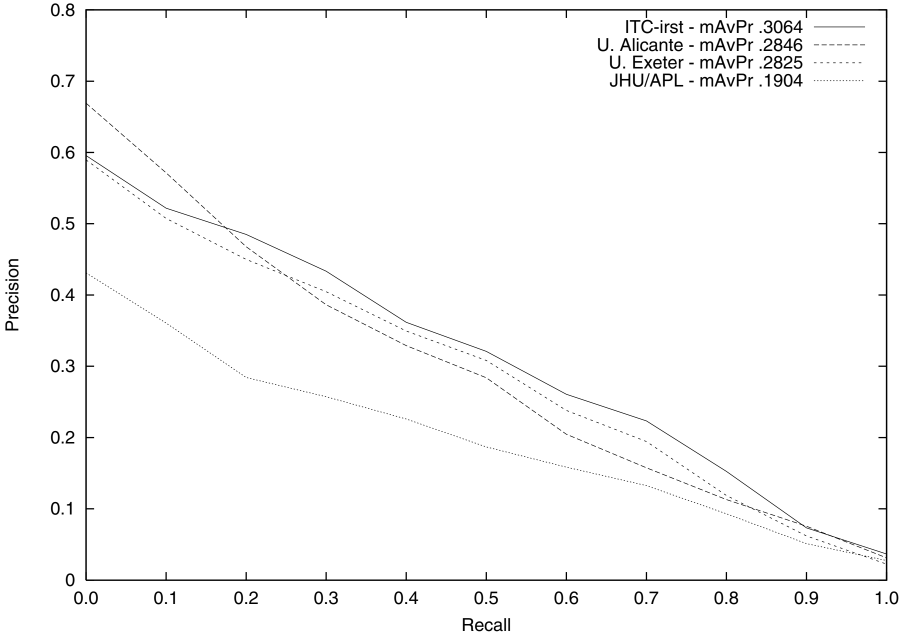

Fig. 2. Precision vs. recall of French-English runs (primary condition)

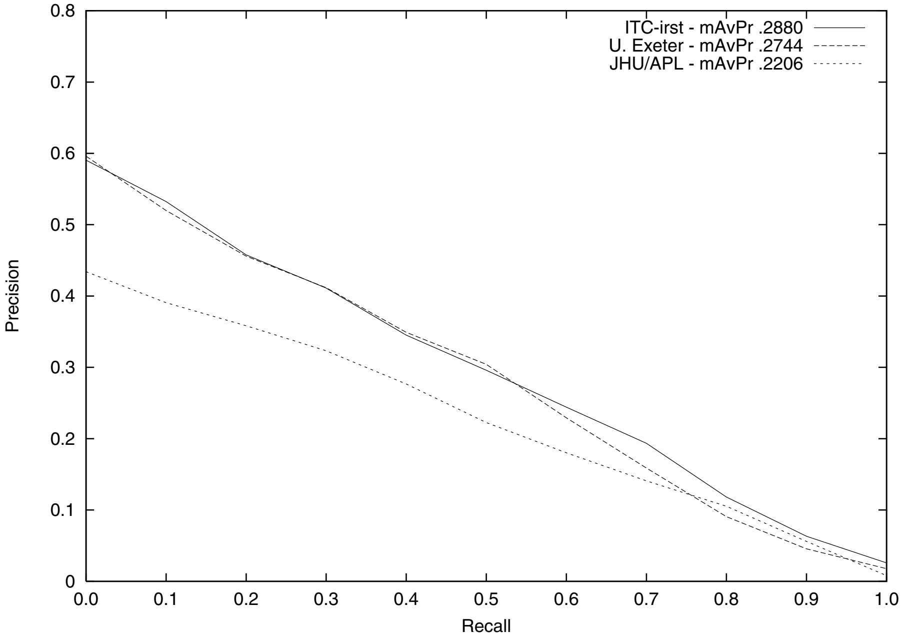

Fig. 3. Precision vs. recall of German-English runs (primary condition)

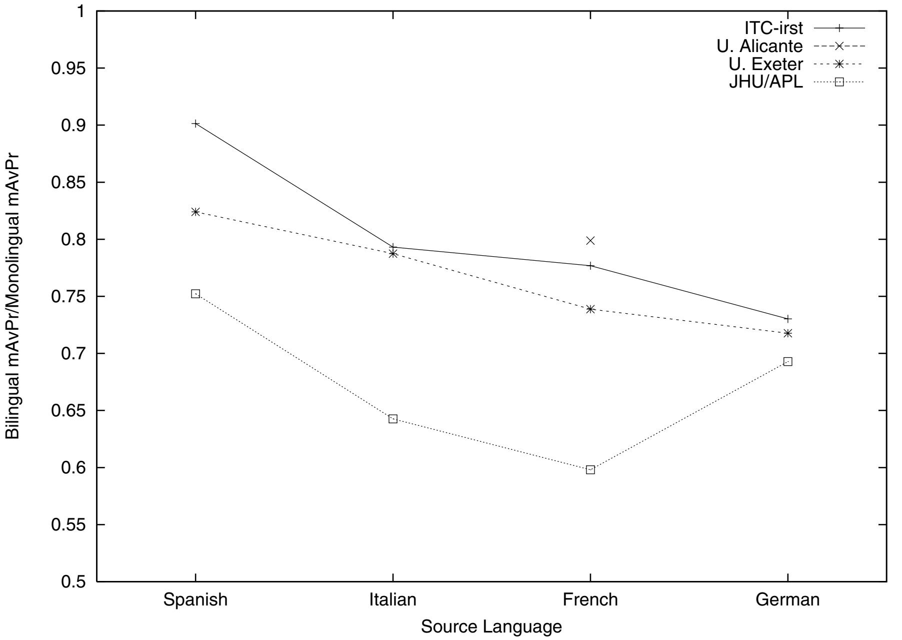

Fig. 4. Ratio between bilingual and monolingual mean-average-pecision for different source languages

Comparing the benefits of using large parallel text collections to improve retrieval performance explored by ITC-irst [1] and U. Exeter [4]. It can be seen that while using a contemporary text collection gives good improvement for ITC-irst, using a text collection from a period several years earlier is not beneficial for U. Exeter. This demonstrates the importance of using suitable text data to help compensate for the errorful transcriptions of the spoken documents.

The JHU submission [6] notes that using word stems performs better for their monolingual retrieval system whereas n-grams are better for bilingual retrieval. This is an interesting result and presumably relates to the translation accuracy and coverage of words vs n-grams for their system.

5 Concluding Remarks

Results from the CLEF 2003 CLSDR task show that as expected bilingual performance is lower for all participants than the comparative English monolingual run. However, the degree of degraded performance is shown to depend on the translation resources used. It has also been shown that different indexing units can be more effective for monolingual and bilingual retrieval on this data set. These are interesting observations and deserve further investigation.

References

1. Bertoldi, N. and Federico, M: ITC-irst at CLEF 2003: Cross-Language Spoken Document Retrieval. In this volume.
2. Garafolo, J. S., Auzanne, C. G. P. and Voorhees, E. M.: The TREC Spoken Document Retrieval Track: A Success Story. In Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access, Paris, 2000, 1–20.
3. Jones, G. J. F. and Federico, M.: 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, 446–457.
4. Jones, G. J. F. and Lam-Adesina, A. M.: Exeter at CLEF 2003: Cross-Language Spoken Document Retrieval Experiments. In this volume.
5. Llopis, F. and Martinez-Barco, P.: Spoken Document Retrieval Experiments with the IR-n system. In this volume.
6. McNamee, P. and Mayfield, J.: N-grams for Translation and Retrieval in CL-SDR. In this volume.