CLEF 2004 Cross-Language Spoken Document Retrieval Track

Marcello Federico[1], Nicola Bertoldi[1], Gina-Anne Levow[2] and
Gareth J.F. Jones[3]

[1] ITC-irst, Italy
[2] University of Chicago, U.S.A.
[3] Dublin City University, Ireland
email: {federico,bertoldi}@itc.it, levow@cs.uchicago.edu,
gareth.jones@computing.dcu.ie

Abstract. This paper summarizes the Cross-Language Spoken Document Retrieval (CL-SDR) track held at CLEF 2004. The CL-SDR task at CLEF 2004 was again based on the TREC-8 and TREC-9 SDR tasks. This year the CL-SDR task was extended to explore the unknown story boundaries condition introduced at TREC. The paper reports results from the participants showing that as expected cross-language results are reduced relative to a monolingual baseline, although the amount to which they are degraded varies for different topic languages.

1 Introduction

The CLEF Cross Language Spoken Document Retrieval (CL-SDR) track aims to evaluate CLIR systems for spoken document collections. The CLEF 2004 CL-SDR track once again takes as its starting point data prepared by NIST for the TREC 8-9 SDR tracks [1]. In particular, the task consists of retrieving news stories within a repository of about 550 hours of American English news. The original English short search topics were manually formulated in other languages, e.g. French or German, to form a CL-SDR task. Retrieval is performed on automatic transcriptions made available by NIST, and generated using different speech recognition systems.

For CLEF 2004, the CL-SDR task was extended to include the unknown story boundaries condition introduced in the TREC SDR evaluations. Whereas for the previous CL-SDR evaluation [2], the transcription was manually divided into individual story units, participants were this year provided only with the unsegmented transcripts. For each search topic, 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. In this way, relevance is assessed by checking if the provided time index falls inside the manually judged relevant stories. According to the NIST evaluation protocol, systems generating results corresponding to the same stories are penalized. In fact, successive time indexes falling in the same story are marked as non relevant results.

2 Data Specifications

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

Queries are based on the 100 English topics in short format from the TREC 8 and TREC 9 SDR tasks, and the corresponding relevance assessments. For the CLIR task, the topics were translated by native speakers into Dutch, Italian, French, German, and Spanish. The existing SDR retrieval scoring software was used for the known and unknown story boundary conditions.

Of the available 100 topics, the first 50 (topic 074 to topic 123) were designated for system development, and the latter 50 (topic 124 to topic 173) for testing. Submission format and evaluation criteria followed the same conventions as those that were used at the 2000 TREC-9 SDR track[1].

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, any available transcript,with parallel data.

3 Participants

Two sites participated in the evaluation: University of Chicago (USA) and ITC-irst (Italy). A brief description of each system is provided.

3.1 CL-SDR System by University of Chicago

Runs were submitted for both the baseline English monolingual task and the French-English cross-language task, using only the resources provided by CLEF with no external resources.

Topic Processing Topic processing aimed to enhance retrieval of the potentially errorful ASR transcriptions through pseudo-relevance feedback expansion. The baseline conditions required the use of only the CLEF provided resources. This restriction limited our source of relevance feedback to the ASR transcriptions, segmented as described below. For both the monolingual English and the

[1] See `http://www.nist.gov/speech/tests/sdr/sdr2000/sdr2000.htm`.

English translations of the original French topics, we performed the same enrichment process. We employed the INQUERY API to identify enriching terms based on the top 10 ranked retrieved segments and integrated these terms with the original query forms. Our hope was that this enrichment process would capture both additional on-topic terminology as well as ASR-specific transcriptions.

For the French-English cross-language condition, we performed dictionary-based term-by-term translation, as described in [3]. We employed a freely available bilingual term list (`www.freedict.com`). After identifying translatable multi-word units based on greedy longest match in the term list, we used a stemming backoff translation approach with statistically derived stemming rules [4], matching surface forms first and backing off to stemmed form if no surface match was found. All translation alternatives were integrated through structured query formulation [5].

Spoken Document Processing This year the SDR track focused on the processing of news broadcasts with unknown story boundaries. This formulation required that sites perform some automatic segmentation of the full broadcasts into smaller units suitable for retrieval. Using an approach inspired by [6], we performed story segmentation as follows. First we created 30 second segments based on the word recognition time stamps using a 10 second step to create overlapping segment windows. These units were then indexed using the INQUERY retrieval system version 3.1p1 with both stemming and standard stopword removal.

Retrieval Segment Construction To produce suitable retrieval segments, we merged the fine-grained segments returned by the base retrieval process on a per-query basis. For each query, we retrieved 5000 fine-grained segment windows. We then stepped through the ranked retrieval list merging overlapping segments, assigning the rank of the higher ranked segment to the newly merged segment. We cycled through the ranked list until convergence. The top ranked 1000 documents formed the final ranked retrieval results submitted for evaluation.

3.2 CL-SDR System by ITC-irst

The ITC-irst system for the CLEF 2004 CL-SDR task was based on the following three processing steps:

1. A collection of news segments is automatically created from the continuous stream of transcripts. Text segments are produced with a shifting time-window of 30 seconds, moved with steps of 10 seconds. Segments are also truncated if a silence period longer than 5 seconds is found.
2. The resulting overlapping texts are used as the target document collection for our text CLIR system [8].

3. Entries in the ranking list which correspond to overlapping segments are merged.

The implemented method works as follows. All retrieved segments of the same news show are sorted by their start time. The first retrieved segment is assumed as the beginning of a new story. If the second segment overlaps with the first, the two are merged, and the time extent of the current story is adjusted, and so on. If a following segment does not overlap with the current story, the current story is saved in a stack, and a new story begins. Finally, for all stories in the stack, only the segments with the highest retrieval score are considered. The process is repeated for all news show files with at least one entry in the rank list. The resulting list of non overlapping segments is then sorted according to the original retrieval score.

4 Results

Table 1 shows a summary of the participants results. For the primary condition, there is a considerable loss in retrieval effectiveness for cross-language relative to monolingual retrieval. This reduction in average precision varies between about 40% and 60%. These figures are larger than those observed for the known story boundary test condition in the CLEF 2003 CL-SDR task [2]. One possible explanation is the small size of the document segments used for the unknown story boundary condition. The combination of errorfully translated short topic statements with these inaccurately transcribed document segments may be responsible for this effect, where, since both are short, redundancy effects, which often help to compensate for transcription and translation errors in SDR and CLIR respectively, will often be very limited.

Table 1. Mean average precision statistics of submitted runs.

Site	Source	Primary	Secondary
ITC-irst	Monolingual	0.306	0.359
	French	0.182 (-40.5%)	0.233 (-35.1%)
	German	0.158 (-48.4%)	0.205 (-42.9%)
	Italian	–	0.251 (-30.1%)
	Spanish	–	0.299 (-16.7%)
U. Chicago	Monolingual	0.296	–
	French	0.108 (-63.5%)	–

As we would expect based on previous work on SDR [2], the use of additional data resources produces an improvement in absolute retrieval performance figures in all cases, although the relative cross language reduction is still very large for all conditions except for Spanish topic translation.

5 Concluding Remarks

The participation of only two groups in the CL-SDR task at CLEF 2004 was disappointing. The comparative results for monolingual and cross-language retrieval in this paper illustrate that effective CL-SDR is a non-trivial task. Unfortunately, since the TREC SDR task, on which the CL-SDR task was based, has previously been investigated extensively both at TREC and in our own earlier CL-SDR investigations [2], it was probably not a sufficiently exciting challenge to encourage wider participation. However, we still regard CL-SDR as both an interesting research problem and a technology which, as with text CLIR, when sufficiently effective and robust may have significant practical applications.

For CLEF 2005 we plan to introduce a brand new CL-SDR task using on a new document collection taken from an entirely different domain, and using a more challenging set of topic languages. The initial task will be based on English documents, but we expect to extend this to more challenging document languages in future years.

References

[1] 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, pages 1–20, Paris, 2000.

[2] Federico, M. and Jones, G. J. F.: The CLEF 2003 Cross-Language Spoken Document Retrieval Track. In Proceedings of Workshop of the Cross-Language Evaluation Forum (CLEF 2003), Peters, C., et al. editors, pages 646-652, Lecture Notes in Computer Science (LNCS 3237), Springer, Heidelberg, Germany, 2004.

[3] Levow, G.–A., Oard, D. W., and Resnik, P.: Dictionary-Based Techniques for Cross-Language Information Retrieval. Information Processing and Management. In press.

[4] Oard, D. W., Levow, G.–A., and Cabezas, C.: CLEF Experiments at the University of Maryland: Statistical Stemming and Backoff Translation Strategies. In Proceedings of Workshop of the Cross-Language Evaluation Forum (CLEF 2000), Peters, C., editor, pages 176-187, Lecture Notes in Computer Science (LNCS 2069), Springer, Heidelberg, Germany, 2001.

[5] Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 55-63, ACM, 1998.

[6] Abberley, D., Renals, S., Cook, G., and Robinson, T.: Retrieval Of Broadcast News Documents With the THISL System. In Proceedings of the Seventh Text REtrieval Conference (TREC-7), Voorhees, E. M., and Harman, D., editors, pages 181–190, NIST Special Publication 500-242, 1999.

[7] Callan, J. P., Croft, W. B., and Harding, S. M.: The INQUERY Retrieval System. In Proceedings of the Third International Conference on Database and Expert Systems Applications pages 78–83, Spinger Verlag, 1992.

[8] Bertoldi, N., and Federico, M.: Statistical Models for Monolingual and Bilingual Information Retrieval. Information Retrieval, (7):51–70, 2004.