CLEF 2004 Cross-Language Spoken Document Retrieval Track

Marcello Federico[1], Nicola Bertoldi[1], Gina-Anne Levow[2], Gareth J.F. Jones[3]
[1] ITC-irst, Italy
[2] University of Chicago, U.S.A.
[3] Dublin City University, Ireland
{*federico,bertoldi*}*@itc.it, levow@cs.uchicago.edu, gareth.jones@computing.dcu.ie*

Abstract

This is a summary report about the Cross-Language Spoken Document Retrieval Track held at CLEF 2004. The report gives brief details of CL-SDR task based again this year on the TREC 8-9 SDR task. This year the CL-SDR task worked with an unknown story boundaries condition. 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 topic languages.

1 Introduction

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 [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 queries were manually formulated in other languages, 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.

As a difference with respect to last year's task [2], systems could not rely on known story-boundaries within the shows, an unknown story boundary condition. Whereas previously the transcription was manually divided into individual story units, participants were this year provided onto with the unsegmented transcript. Hence, 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. In this way, relevance is assessed by checking if the provided time indexes fall inside the manually judged relevant stories. According to the NIST evaluation protocol, systems generating results corresponding to the very same stories are penalized. In fact, successive time indexes falling in the same story are marked as non relevant results.

2 Data Specifications

The target 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 are allowed to use parallel document collections available through the Linguistic Data Consortium.

Queries are based on a collection of 100 English topics in short format for which relevance assessments are available. For the sake of CLIR, queries were translated by native speakers into Dutch, Italian, French, German, and Spanish. Retrieval scoring software is available both for the known and unknown story boundary conditions.

Of the available 100 topics, the first 50 (topic 074 to topic 123) were intended for system development, while the latter 50 (topic 124 to topic 173) for testing. Submission format and evaluation criteria followed the same conventions as 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, all transcripts,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 U. Chicago

The University of Chicago participated in the CLEF 2004 spoken document retrieval task. 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.

3.1.1 Query Processing

Query 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 English translations of the original French queries, 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].

3.1.2 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.

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

3.1.3 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 is based on the following three processing steps.

First, 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. Moreover, segments are also truncated if a silence period longer than 5 seconds is found.

Second, the resulting overlapping texts are used as target document collection by means of a text CLIR system [8].

Third, entries in the ranking list which correspond to overlapping segments are properly 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

Site	Source	Primary	Secondary
ITC-irst	Monolingual	.3059	.3586
	French	.1816	.2330
	German	.1584	.2051
	Italian		.2510
	Spanish		.2990
U. Chicago	Monolingual	.2963	-
	French	.1084	-

Table 1: Mean average precision statistics of submitted runs.

The results in Table 1 show that, particularly in 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 errorfully short topic statements with the inaccurately transcribed document segments may be responsible for this effect.

As we would expect 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.

References

[1] 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.

[2] M. Federico and G. J. F. Jones. The CLEF 2003 Cross-Language Spoken Document Retrieval Track. In *Proceedings of Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, Trondheim, Norway, C. Peters et al. editors, Springer-Verlag, 2004.

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

[4] D. W. Oard,G.–A. Levow and C. Cabezas, 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)*, Lisbon, Portugal, C. Peters editor, pages 176-187, Springer-Verlag, 2001.

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

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

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