Out-of-the-box strategy for Rich Speech Retrieval @ MediaEval 2011

Wouter Alink
Spinque
Utrecht, The Netherlands
wouter@spinque.com

Roberto Cornacchia
Spinque
Utrecht, The Netherlands
roberto@spinque.com

ABSTRACT

Evaluation tracks offer valuable opportunities to measure scientific and technological advances. Spinque approaches challenges as the MediaEval Rich Speech Recognition task with the additional goal of developing solutions that can easily transferred from academic labs to industry. The system used during this evaluation was obtained with minimal effort and no manual optimisation and yet it provides a reasonably good baseline to improve upon. More importantly, it is by nature an extensible approach, based on the concept of declarative search strategies, rather than an ad-hoc search system.

1. INTRODUCTION

Our participation in the MediaEval Rich Speech Recognition task, described in [3], has been inspired by the quest for finding a simple, fast, robust, and effective approach to searching in speech transcripts. We used our generic search framework to instantiate a specific search solution for this task, with the explicit goal of producing reasonable results in the space of a few hours, including index creation, search strategy modelling and evaluation. As for example argued in [2], standard textual IR techniques can be applied to speech transcripts, even when the transcripts are not perfect. Our runs focus on textual search with different query keyword combinations and with rank refinement at different levels of retrieval unit granularity.

2. SPINQUE FRAMEWORK

We modelled and executed our runs as *search strategies* within the Spinque framework. This is a prototype environment where search processes are divided into two phases: the search strategy definition and the actual search.

Modelling search strategies in this framework corresponds to designing graph structures, where edges represent dataflows consisting of terms, documents (e.g. speech-transcripts), and document-sections. The nodes connected by such edges are pre-defined, general-purpose operational blocks, that either provide source data (the speech transcripts and the topics) or modify their input data-flow applying operations such as extraction of specific sections from documents or ranking of sections and documents, to name a few.

Search strategies defined in this framework are automatically translated into a probabilistic relational query language and executed on top of an SQL database engine.

The same framework has also been used to participate in other evaluation tracks, such as CLEF-IP [1].

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

3. DESCRIPTION

The speech transcripts were indexed at two levels of granularity: as whole documents as well as individual *SpeechSegment* sections. We did not use the tags and the video keyframes provided, nor any other source of evidence.

Our runs can be described as follows:

run1 First, all words from *title* (weight 0.2) and all words from *short-title* (weight 0.8) are used to search all documents in the collection. Then, all the SpeechSegment sections within those documents are searched using the same keywords. The start of the section is returned as the result. This strategy is depicted in Figure 1.

run2 the same as run1, except that all terms from *title* get a weight of 0.0 and all terms from *short-title* get a weight of 1.0. This basically discards the terms from *title*.

run3 the same as run1, except that all terms from *title* get a weight of 1.0 and all terms from *short-title* get a weight of 0.0. This basically discards the terms from *short-title*. Run3 should be considered as the “required run”.

Textual ranking is performed with the BM25 [4] retrieval method, with standard parameters $b = 0.75$ and $k_1 = 1.2$. The weights 0.2 (words from *title*) and 0.8 (words from *short-title*) have been found as the local optimum using a hill climbing approach.

4. RESULTS AND FINDINGS

The average time for retrieving results for a topic was 230ms. This time includes “compiling” the search strategy (i.e. translating it into SQL queries) out-of-the-box and without manual optimisations, and the overhead for generating the run-files. A glitch later found in our indexer may have altered results marginally: a few documents have not been included in our index and therefore not retrieved.

The evaluation scores for the 3 submitted runs are shown in Table 1. Scores have been measured with window sizes of 10, 30, and 60 seconds. Overall scores are reasonably satisfying for a simple keyword-search approach. As expected, the combination of both the *title* and the *short-title* yield a better result than the individual runs. Best results were

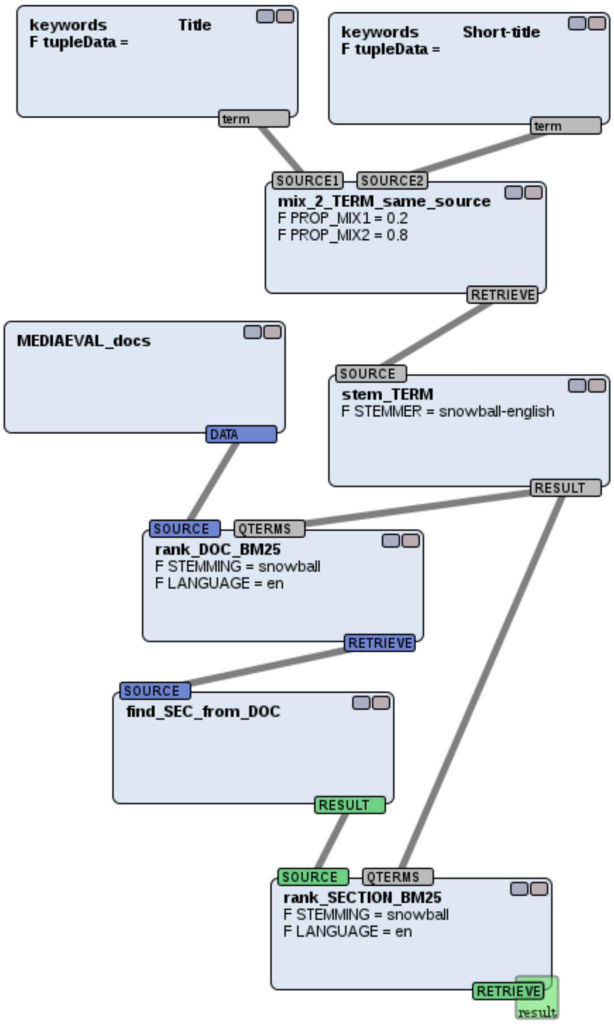

Figure 1: Search strategy using both title and short-title as input, first searching the whole transcript documents, then refining into sections.

	Weights for		Window size (seconds)		
	title	short-title	10	30	60
Run 1	0.2	0.8	0.1320	0.2210	0.2724
Run 2	0.0	1.0	0.1164	0.1816	0.2231
Run 3	1.0	0.0	0.1054	0.1630	0.1968

Table 1: mGAP scores for the runs on the test-set with 50 topics (step size is 10 seconds)

found on the test-set assigning a larger weight to *short-title* keywords, which suggests that full titles may carry off-topic words which yield lower precision.

We found that searching short sections produced disappointing rankings, probably due to a non fine-tuned document-length normalisation. Both parameter configurations used (for BM25 and for the title / short-title keyword mixture) could be improved with a more exhaustive exploration of their search space. The simplicity of the strategies used and the small size of the corpus at hand would make this approach feasible indeed, which is not the case in general.

One more direction for possible improvements is to experiment with a more fine-grained zooming in, with search windows of e.g. entire documents followed by 10 minute, 1 minute and 5 seconds speeches. Such a multi-stage strategy would likely retain recall and improve precision at every iteration.

5. CONCLUSIONS

The main contribution of this paper is to show how a specific search engine for speech transcripts of reasonable quality can be instantiated with minimal effort. While out-of-the-box text search is not unique to Spinque's framework, the ability to play with retrieval units of different granularities and combine query and/or data sources easily is not common.

We plan to improve on our first speech retrieval evaluation in two ways: firstly, by automating as much as possible the optimisation of search strategies' free parameters, including the choice of unit retrieval granularities; secondly, by building on top of this optimised baseline with the addition of more sources of evidence that may be available (such as tags and video material).

6. REFERENCES

[1] W. Alink, R. Cornacchia, and A.P. de Vries. Searching clef-ip by strategy. In *CLEF 2009, Revised Selected Papers, Part I*. Springer, 2010.

[2] James Allan. Perspectives on information retrieval and speech. In *Information Retrieval Techniques for Speech Applications*, volume 2273 of *Lecture Notes in Computer Science*, pages 323–326. Springer Berlin / Heidelberg, 2002.

[3] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G.J.F. Jones. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.

[4] S.E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Third Text REtrieval Conference (TREC 1994)*, 1994.