New Challenges for Cross-Language Information Retrieval: multimedia data and the user experience

Gareth Jones

Department of Computer Science

University of Exeter, U.K.

Overview

- Introduction
- Review of Spoken Document Retrieval
- Cross-Language Speech Retrieval
- Evaluation for CLIR, SDR & CLSR

Introduction

- CLIR research has focussed primarily on text retrieval.
- Documents may originate in different media: typed text, spoken data, document images - OCR (paper, video).
- Requirement for systems capable of effective cross-language retrieval from mixed-media collections.
- CLIR and SDR evaluations have focussed on document retrieval.
  - An important issue is how can the user most effectively access information from within these retrieved documents.

Spoken Document Retrieval

- First work in the early 1990's: (Rose 91) (Glavitsch and Schauble 92)
- Further studies: Cambridge University VMR, CMU Informedia, ETH.
- TREC:
  - TREC-6: known item search
  - TREC-7: more data, adhoc task.
  - TREC-8: more data.
  - TREC-9: unknown story boundaries, etc.

Spoken Document Retrieval

- Most TREC participants index data using Large Vocabulary Recognition (LVR).
- Retrieval performance similar to text using *document expansion* (AT&T).
- SDR track is finished!

Spoken Document Retrieval - New Challenges!

- More languages - European, Asian
- Information access from retrieved documents.
  - how can the user most efficiently and/or effectively access the information they need?
  - does the document transcription satisfy the user information need?
  - interactive testing of multi-modal interfaces?
- Spoken search requests.

Cross-Language Speech Retrieval

- Retrieve spoken documents across languages.
- Existing studies:
  - ETH (1997) - French requests to retrieve German documents.
  - Exeter (2000) - French requests to retrieve English documents.
- small scale studies indicate that:
  - CLIR techniques, e.g. pseudo relevance feedback are effective for CLSR.
  - performance retrieval degradations from CLIR and SDR are additive.

Evaluation

- Recall and Precision:
  - tell us whether we have retrieved the relevant documents and with what accuracy.
  - they don't tell us whether the user can identify relevant documents or extract information from within them.
- CLIR: can the user read the document language?
  - graphical document representations: TileBars, ThumbNails.
  - gloss translations.
- Investigate efficiency and accuracy issues.

Evaluation

- CLSR:
  - need for standard test collections.
  - has all the above information access issues and more!
- How should users assess relevance and access information from within individual retrieved documents?