Cross-Language Information Retrieval

Carol Peters
ISTI-CNR, Pisa

Cross-Language Information Retrieval (CLIR)

State-of-the-Art

- The basic cross-language text retrieval system technology now exists for bilingual and multilingual retrieval
- System performance is ca. 90% of monolingual

But gap between R&D results and application requirements

- Large Web search engines do not offer CLIR
- Few commercial information services offer CLIR

Current systems don’t meet needs of generic user

CLIR: Ultimate Goal

Fully multilingual, multimodal information retrieval systems

- capable of processing a query in any medium and any language
- finding relevant information from a multilingual multimedia collection containing documents in any language and form,
- and presenting it in the style most likely to be useful to the user

CLIR: Subgoals

- Fully multilingual test retrieval
- Cross-language multimodal systems
- Multilingual question answering
- Cross-language interactive systems

Multilingual Text Retrieval (MLTR) ETD: 2007

Goal: Truly multilingual systems - L1 -> Ln

Work needed on:

- Most appropriate system architecture
  - Translation + IR for each language or unified framework
- Overcoming the translation bottleneck
  - Improve LRs/optimise pivot language approaches/conceptual interlingua/language independent methods

2007: MLTR systems capable of including any new language within 1 month

Cross-Language Multimodal Systems ETD: 2007 for 1st results

Goal: successful retrieval across languages in collections of multimedia (video/image/speech/text) – combination of language dependent and language independent methods:

- **2 Stages:**
  - ➢ C-L retrieval as particular form of text retrieval
    e.g. image captions, noisy speech transcriptions
  - ➢ C-L retrieval as combination of media-specific methods
    e.g. text-based and content-based methods

2007: Testing on target collections of multimedia data in five languages

Multilingual Question Answering ETD: 2007 for 1st results

Goal: CLIR systems capable of precise IE

2 stages:

- development of monolingual **non-English** systems
- development of C-L QA systems

(combination of NLP and IR tools)

2007: Testing on target collections in five languages

Interactive Cross-Language Systems ETD: 2007

Goal: Systems that help user in query formulation and results selection and interpretation

2007: On-line multilingual text retrieval system searching on collections in at least five language with functionality for user-assisted query formulation, refinement, document selection and interpretation.

Recommendations

- Internationally funded research programme that prepares a roadmap and promotes its completion through the orgnisation of evaluation campaigns with appropriate tasks designed to stimulate system development in the directions identified