LBSC 796/INFM 718R: Week 11

Cross-Language and Multimedia Information Retrieval

Jimmy Lin
College of Information Studies
University of Maryland

Monday, April 17, 2006

Topics covered so far…

- Evaluation of IR systems
- Inner workings of IR black boxes
- Interacting with retrieval systems
- Interfaces in support of retrieval

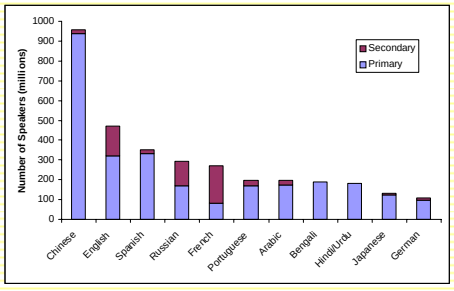
Questions for Today

- What if the collection contains documents in a foreign language?
- What if the collection isn't even comprised of textual documents?

Cross-Language IR

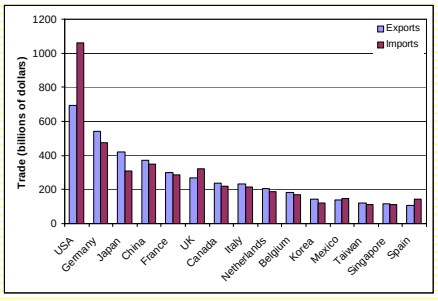
- Or "finding documents in languages you can't read"
- Why would you want to do it?
- How would you do it?

Most Widely-Spoken Languages



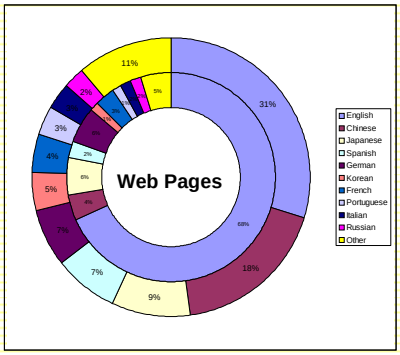
Source: Ethnologue (SIL), 1999

Global Trade



Source: World Trade Organization 2000 Annual Report

Global Internet Users

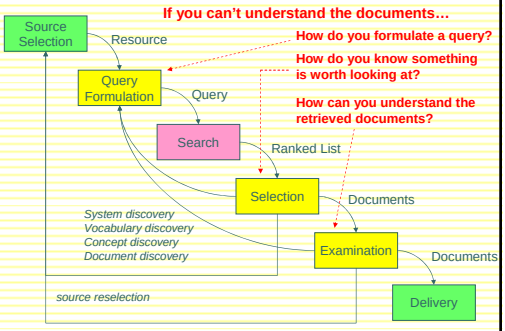


Native speakers, Global Reach projection for 2004 (as of Sept, 2003)

A Community: CLEF

- CLEF = "Cross-Language Evaluation Forum"
- 8 tracks at CLEF 2005
  - Multilingual information retrieval
  - Cross-language information retrieval
  - Interactive cross-language information retrieval
  - Multiple language question answering
  - Cross-language retrieval on image collections
  - Cross-language spoken document retrieval
  - Multilingual Web retrieval
  - Cross-language geographic retrieval

The Information Retrieval Cycle



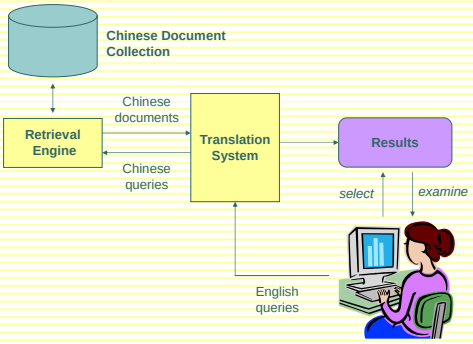
CLIR

- CLIR = "Cross Language Information Retrieval"
- Typical setup
  - User speaks only English
  - Wants access to documents in a foreign language (e.g., Chinese or Arabic)
- Requirements
  - User needs to understand retrieved documents!
  - Interface must support browsing of documents in foreign languages
- How do we do it?

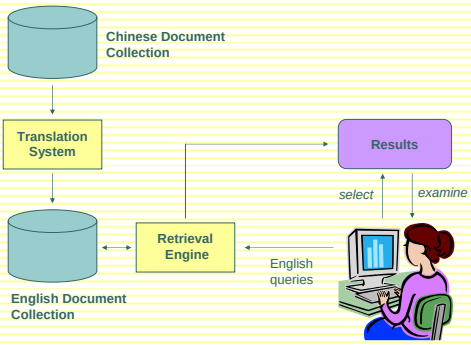
Two Approaches

- Query translation
  - Translate English query into Chinese query
  - Search Chinese document collection
  - Translate retrieved results back into English
- Document translation
  - Translate entire document collection into English
  - Search collection in English
- Translate both?

Query Translation



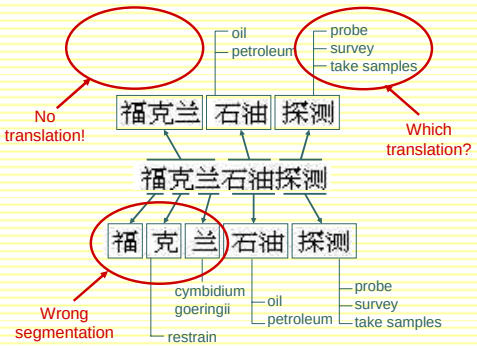
Document Translation



Tradeoffs

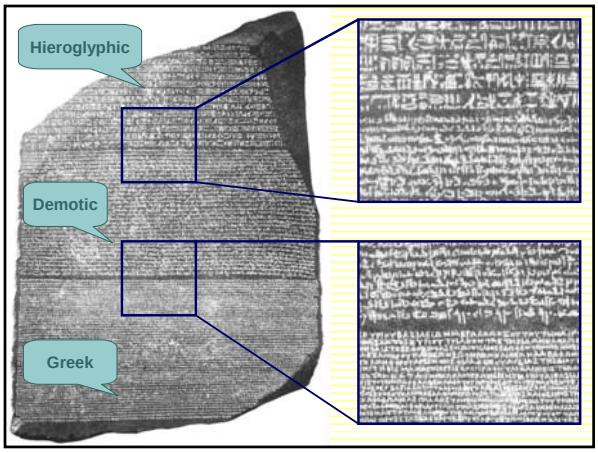
- Query Translation
  - Often easier
  - Disambiguation of query terms may be difficult with short queries
  - Translation of documents must be performed at query time
- Document Translation
  - Documents can be translate and stored offline
  - Automatic translation can be slow
- Which is better?
  - Often depends on the availability of language-specific resources (e.g., morphological analyzers)
  - Both approaches present challenges for interaction

CLIR Issues



Learning to Translate

- Lexicons
  - Phrase books, bilingual dictionaries, …
- Large text collections
  - Translations (“parallel”)
  - Similar topics (“comparable”)
- People



Modern Rosetta Stones

- Newswire:
  - DE-News (German-English)
  - Hong-Kong News, Xinhua News (Chinese-English)
- Government:
  - Canadian-Hansards (French-English)
  - Europarl (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portugese, Spanish, Swedish)
  - UN Treaties (Russian, English, Arabic, …)
- The Bible (many, many languages)

Parallel Corpus

- Example from DE-News (8/1/1996)

English: Diverging opinions about planned tax reform

German: Unterschiedliche Meinungen zur geplanten Steuerreform

English: The discussion around the envisaged major tax reform continues .

German: Die Diskussion um die vorgesehene grosse Steuerreform dauert an .

English: The FDP economics expert , Graf Lambsdorff , today came out in favor of advancing the enactment of significant parts of the overhaul , currently planned for 1999 .

German: Der FDP - Wirtschaftsexperte Graf Lambsdorff sprach sich heute dafuer aus , wesentliche Teile der fuer 1999 geplanten Reform vorzuziehen .

Word-Level Alignment

English
Diverging opinions about planned tax reform

Unterschiedliche Meinungen zur geplanten Steuerreform
German

English
Madam President , I had asked the administration …

Señora Presidenta, había pedido a la administración del Parlamento …
Spanish

Learning Translations

- From alignments, automatically induce a translation lexicon

survey

探测	(p = 0.4)
试探	(p = 0.3)
测量	(p = 0.25)
样品	(p = 0.05)

Multiple Translations **Translation Probabilities**

A Translation Model

- From word-aligned bilingual text, we induce a translation model

$p(f_i \mid e)$ where, $\sum_{f_i} p(f_i \mid e) = 1$

- Example:

p(探测|survey) = 0.4
p(试探|survey) = 0.3
p(测量|survey) = 0.25
p(样品|survey) = 0.05

Using Multiple Translations

- Weighted Structured Query Translation
  - Takes advantage of multiple translations and translation probabilities
- *TF* and *DF* of query term *e* are computed using *TF* and *DF* of its translations:

$$TF(e, D_k) = \sum_{f_i} p(f_i \mid e) \times TF(f_i, D_k)$$

$$DF(e) = \sum_{f_i} p(f_i \mid e) \times DF(f_i)$$

Experiment Setup

- Does weighted structured query translation work?
- Test collection (from CLEF 2000-2003)
  - ~ 44,000 documents in French
  - 153 topics in English (and French, for comparison)
- IR system: Okapi weights
- Translation resources
  - Europarl parallel corpus: ~ 100M on each side
  - GIZA++ Statistical MT toolkit

Does it work?

- Runs:
  - Monolingual baseline
  - One-best translation baseline
  - Weighted structured query translation
- Results:
  - Weighted structured query translation always beats one-best translation
  - Weighted structured query translation performance approaches monolingual performance

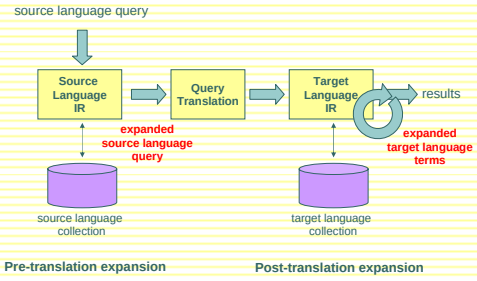
Morphology and Segmentation

- For the query translation approach
  - The retrieval engine needs to perform monolingual IR in a foreign language
  - Morphology and segmentation pose problems
- Good segmenters and morphological analyzers are expensive to develop
- N-gram indexing provides a good solution
  - Use character n-grams based on length of average word
  - Performs about as well as with a good segmenter

Blind Relevance Feedback

- Augment the query representation with related terms
- Multiple opportunities for expansion
  - Before doc translation: Enrich the vocabulary
  - After doc translation: Mitigate translation errors
  - Before query translation: Improve the query
  - After query translation: Mitigate translation errors

Query Expansion/Translation

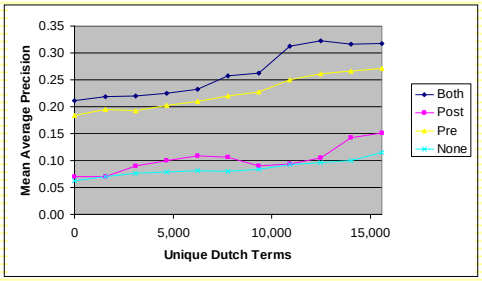


McNamee and Mayfield

- Research questions:
  - What are the effects of pre- and post- translation query expansion in CLIR?
  - How is performance affected by quality of resources?
  - Is CLIR simply measuring translation performance?
- Setup:
  - CLEF 2001 test collection
  - Dutch, French, German, Italian, Spanish queries
  - English documents
  - Varied the size translation lexicons (randomly threw out entries)

Paul McNamee and James Mayfield. (2002) Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. *Proceedings of SIGIR 2002.*

Query Expansion Effect



Lessons

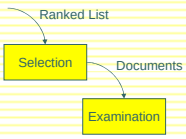
- Both pre- and post- translation expansions help
- Pre-translation expansion is a bigger win… why?
- Translation resources are important!

Interaction

- CLIR poses some unique challenges for interaction
  - How do you help users select translated query terms?
  - How do you help users select document terms for query refinement?
  - How do you compensate for poor translation quality?

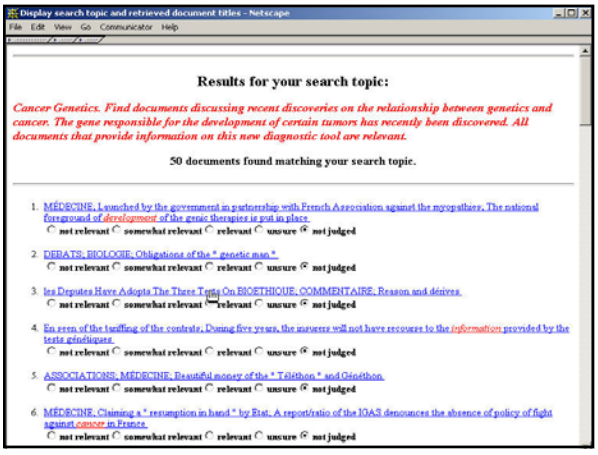
Document Selection

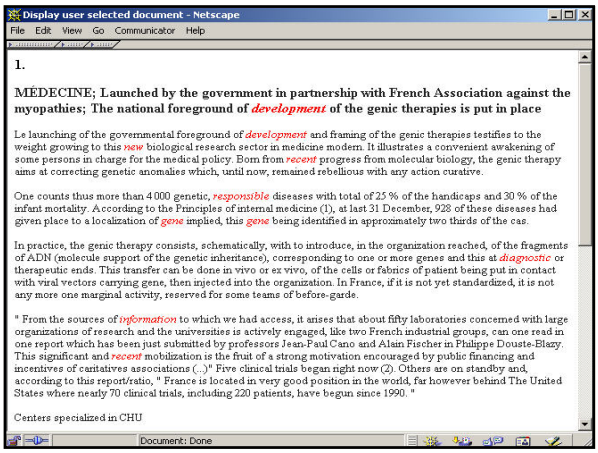
- Can users recognize relevant documents in a cross-language retrieval setting?
- What's the impact of translation quality?



Selection Experiment

- Experimental setup (UMD, iCLEF 2001):
  - English topics, French documents
  - Each user works with the same hit list
  - Can users make relevance judgments?
  - What's the effect of translation quality?
- Comparison of two translation methods:
  - Term-for-term gloss translation (Gloss)
    - Easily built for a wide range of language pairs
    - Widely available bilingual word lists
  - Machine translation (MT)
    - Syntactic/semantic constraints improve accuracy & fluency
    - Used Systran, a commercially available MT system
    - Developing new language pairs is expensive (years)



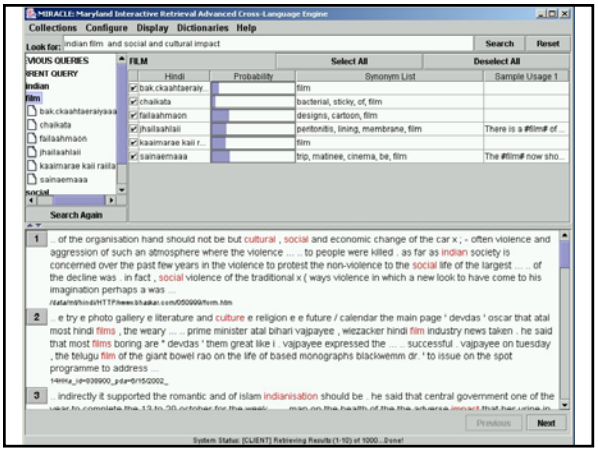


Results

- Quantitative measures:
  - Users with the MT system achieved higher F-score
- Observed behavior (from observational notes):
  - Documents were usually examined in rank order
  - Title alone was seldom used to judge documents as "relevant"
- Subjective reactions (from questionnaires):
  - Everyone liked MT
  - Only one participant liked anything about gloss translation
  - MT was preferred overall

Making MIRACLEs

- Putting everything together in an interactive, cross-language retrieval system…



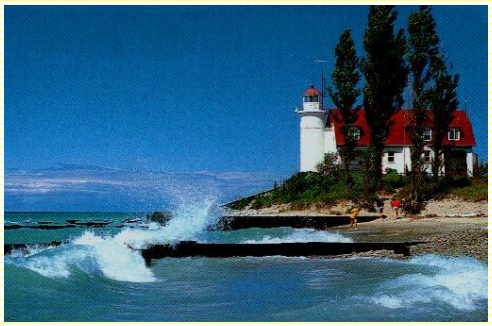
Key Points

- Good translation is the key to cross-language information retrieval
  - Where does one obtain them? (e.g., bilingual dictionaries, aligned text, etc.)
  - How does one use them? (e.g., query translation, document translation, etc.)
- CLIR performance approaches monolingual IR performance
- CLIR presents addition challenges for interaction support

Multimedia Retrieval

- We're primarily going to focus on image and video search

A Picture…



… is comprised of pixels

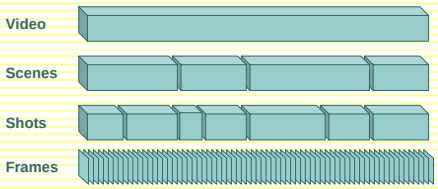
This is nothing new!

Seurat, Georges, A Sunday Afternoon on the Island of La Grande Jatte

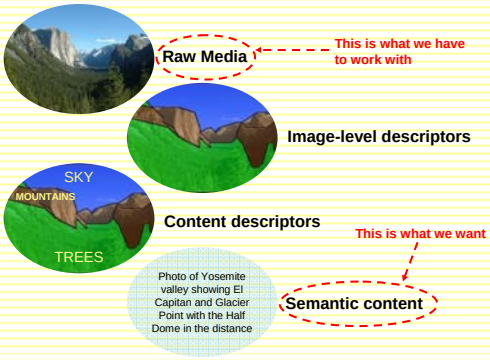
Images and Video

- A digital image = a collection of pixels
  - Each pixel has a “color”
- Different types of pixels
  - Binary (1 bit): black/white
  - Grayscale (8 bits)
  - Color (3 colors, 8 bits each): red, green, blue
- A video is simply lots of images in rapid sequence
  - Each image is called a frame
  - Smooth motion requires about 24 frames/sec
- Compression is the key!

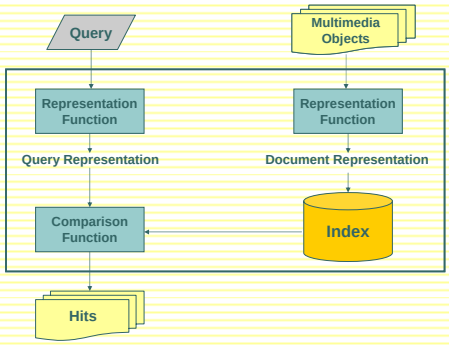
The Structure of Video



The Semantic Gap



The IR Black Box



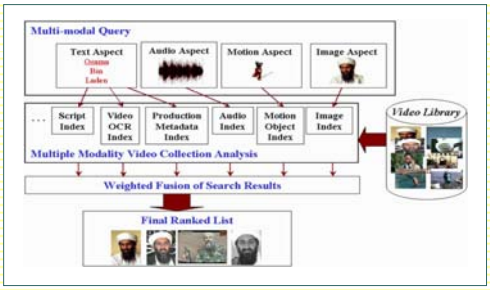
Recipe for Multimedia Retrieval

- Extract features
  - Low-level features: blobs, textures, color histograms
  - Textual annotations: captions, ASR, video OCR, human labels
- Match features
  - From "bag of words" to "bag of features"

Demos

- Google Image Search
  http://images.google.com/
- Hermitage Museum
  http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English
- IBM's MARVEL System
  http://mp7.watson.ibm.com/

Combination of Evidence



TREC For Video Retrieval?

- TREC Video Track (TRECVID)
  - Started in 2001
  - Goal is to investigate content-based retrieval from digital video
  - Focus on the shot as the unit of information retrieval (why?)

  http://www-nlpir.nist.gov/projects/trecvid/
- Test Data Collection in 2004:
  - 74 hours of CNN Headline News, ABC World News Tonight, C-SPAN

Searching Performance

Modality	MAP
Baseline: ASR + Closed Captions (CC)	0.155
ASR + CC + Video OCR	0.177
ASR + CC + VOCR + Image Similarity weighted by query type	0.198
ASR + CC + VOCR + Image Similarity weighted by development set query results	0.207
ASR + CC + VOCR + Image Similarity weighted by development set query results + Person X retrieval	0.218

A. Hauptmann and M. Christel. (2004) Successful Approaches in the TREC Video Retrieval Evaluations. Proceedings of ACM Multimedia 2004.

Interaction in Video Retrieval

- Discussion point: What unique challenges does video retrieval present for *interactive* systems?

Take-Away Message

- Multimedia IR systems build on the same basic set of tools as textual IR systems
  - If you have a hammer, everything becomes a nail
- The feature set is different… but the ideas are the same
- Text is important!