SIGIR-97 Tutorial

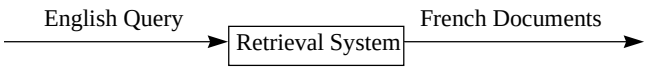
Cross-Language Information Retrieval

Douglas W. Oard
College of Library and Information Services
University of Maryland, College Park
oard@glue.umd.edu

July 27, 1997

Cross-Language IR

- Given a query expressed in one language
- Find info that may be expressed in another
  - Electronic texts
  - Document images
  - Recorded speech [101]
  - Sign language



2

Why Do Cross-Language IR?

- When users can read several languages
  - Eliminates multiple queries
  - Query in most fluent language
- Monolingual users can also benefit
  - If translations can be provided
  - If it suffices to know that a document exists
  - If text captions are used to search for images

3

What We Know

- Dictionaries are very useful
  - Easily get to 50% of monolingual IR effectiveness
  - We can get to about 75% using:
    - Part-of-speech tags
    - Pseudo-relevance feedback
    - Phrase indexing
- Multilingual training corpora are also useful
  - When the corpus is from the right domain

4

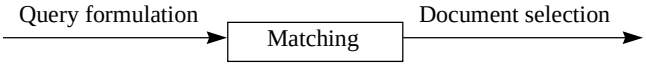
A Little Vocabulary

- Cross-language
  - Cross-lingual, cross-linguistic, translingual
- Multilingual
  - Used to describe cross-language systems
    - Query in English finds documents in English or French
    - Query in French finds documents in English or French
  - Also used for paired monolingual systems
    - Queries in English find documents in English
    - Queries in French find documents in French

5

Scope

- Query formulation
  - Natural language and structured queries
- Matching queries to documents
  - Exact match and ranked retrieval
- Document selection
  - Title translation and transliteration
  - Cross-language gisting



6

Related Issues

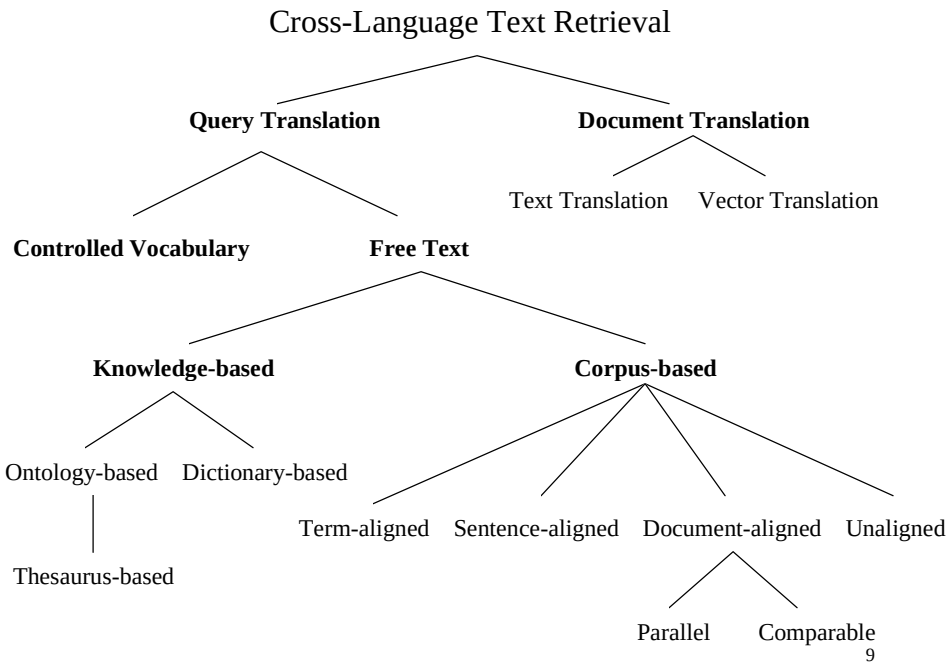
- Multiscript text processing [12]
  - Character sets, writing system, direction, ...
- Language identification [109]
  - Markup, detection
- Language-specific processing [103]
  - Stemming, morphological roots, compounds, …
- Document translation [51]

7

Design Decisions

- What to index?
  - Free text or controlled vocabulary
- What to translate?
  - Queries or documents
- Where to get translation knowledge?
  - Dictionary, ontology, training corpus

8



9

Agenda

- Historical overview
  - Controlled vocabulary techniques
- Knowledge-based techniques
- Corpus-based techniques
- Performance evaluation
- System integration
- Research directions

10

Cross-Language Information Retrieval

Historical Overview
and
Controlled Vocabulary Techniques

Early Development

- 1964 International Road Research [81]
  - English, French and German thesaurus
- 1969 Pevzner [78]
  - Exact match with a large Russian/English thesaurus
- 1970 Salton [94]
  - Ranked retrieval with small English/German dictionary
- 1971 UNESCO [107]
  - Proposed standard for multilingual thesauri

12

Controlled Vocabulary Matures

- 1977 IBM STAIRS-TLS [95]
  - Large-scale commercial cross-language IR
- 1978 ISO Standard 5964
  - Guidelines for developing multilingual thesauri
- 1984 EUROVOC thesaurus [34]
  - Now includes all 9 EC languages
- 1985 ISO Standard 5964 revised [46]

13

Free Text Developments

- 1970, 1973 Salton
  - Hand coded bilingual dictionaries
- 1990 Latent Semantic Indexing [53]
  - French/English using Hansard training corpus
- 1994 European multilingual IR project [84]
  - Medium-scale recall/precision evaluation
- 1996 SIGIR Cross-lingual IR workshop
  - And over 10 conferences and workshops since!

14

How Controlled Vocabulary Works

- Thesaurus design [102]
  - Design a knowledge structure for domain
  - Assign a unique “descriptor” to each concept
    - Include “scope notes” and “lead-in vocabulary”
- Document indexing
  - Read the document, assign appropriate descriptors
- Retrieval
  - Select desired descriptors, use exact match retrieval

15

Multilingual Thesauri

- Adapt the knowledge structure
  - Cultural differences influence indexing choices
- Use language-independent descriptors
  - Matched to a unique term in each language
- Three construction techniques [46]
  - Build it from scratch
  - Translate an existing thesaurus
  - Merge monolingual thesauri

16

Advantages over Free Text

- High-quality concept-based indexing
  - Descriptors need not appear in the document
- Knowledge-guided searching
  - Good thesauri capture expert domain knowledge
- Excellent cross-language effectiveness
  - Up to 100% of monolingual effectiveness
- Understandable retrieval results
- Efficient implementation

17

Limitations

- Costly to create
  - Design knowledge structure, index each document
- Costly to maintain
  - Document indexing, vocabulary and concept change
- Hard to use
  - Vocabulary choice, knowledge structure navigation
- Limited scope
  - Domain must be chosen at design time

18

Addressing the Limitations

- Automatic thesaurus construction [39]
  - Exploit large multilingual training corpora
  - Thesaurus merging tools
- Thesaurus maintenance tools
- Machine aided indexing
- Graphical search interfaces
  - Visualize knowledge structure

19

Current Research and Practice

- Unified Medical Language System
  - Integrating medical coverage of many thesauri
- Access Innovations [41]
  - Machine aided indexing
- University of Huddersfield [83]
  - Graphical search interface using EUROVOC
- VTLS
  - Library catalog with cross-language subject search

20

Cross-Language Information Retrieval

Knowledge-based Techniques
for Free Text Searching

Knowledge Structures for IR

- Ontology
  - Representation of concepts and relationships
- Thesaurus
  - Ontology specialized for retrieval
- Bilingual lexicon
  - Ontology specialized for machine translation
- Bilingual dictionary
  - Ontology specialized for human translation

22

Query vs. Document Translation

- Query translation
  - Very efficient for short queries
    - Not as big an advantage for relevance feedback
  - Hard to resolve ambiguous query terms
- Document translation
  - May be needed by the selection interface
    - And supports adaptive filtering well
  - Slow, but only need to do it once per document
    - Poor scale-up to large numbers of languages

23

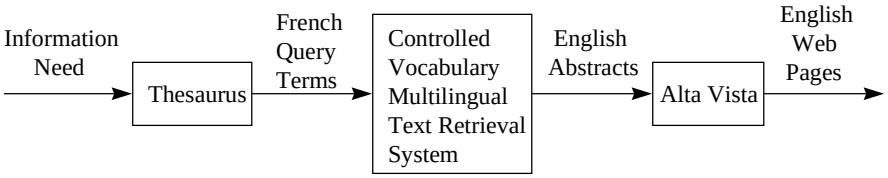
Document Translation Example

- Approach
  - Select a single query language
  - Translate every document into that language
  - Perform monolingual retrieval
- Long documents provide enough context
  - And many translation errors do not hurt retrieval
- Much of the generation effort is wasted
  - And choosing a single translation can hurt

24

Query Translation Example

- Select controlled vocabulary search terms
- Retrieve documents in desired language
- Form monolingual query from the documents
- Perform a monolingual free text search



25

Machine Readable Dictionaries

- Based on printed bilingual dictionaries
  - Becoming widely available
- Used to produce bilingual term lists
  - Cross-language term mappings are accessible
    - Sometimes listed in order of most common usage
  - Some knowledge structure is also present
    - Hard to extract and represent automatically
- The challenge is to pick the right translation

26

Unconstrained Query Translation

- Replace each word with every translation
  - Typically 5-10 translations per word
- About 50% of monolingual effectiveness
  - Main problem is ambiguity
  - Example: Fly (English)
    - 8 word senses (e.g., to fly a flag)
    - 13 Spanish translations (enarbolar, ondear, …)
    - 38 English retranslations (hoist, brandish, lift…)

27

Exploiting Part-of-Speech Tags

- Constrain translations by part of speech [15]
  - Noun, verb, adjective, …
  - Effective taggers are available
- Works well when queries are full sentences
  - Short queries provide little basis for tagging
  - Constrained matching can hurt monolingual IR
    - e.g., nouns in queries can match verbs in documents

28

Phrase Indexing

- Improves retrieval effectiveness two ways
  - Phrases are less ambiguous than single words
  - Idiomatic phrases translate as a single concept
- Three ways to identify phrases
  - Semantic (e.g., appears in a dictionary)
  - Syntactic (e.g., parse as a noun phrase)
  - Cooccurrence (words found together often)
- Semantic phrase results are impressive [43]

29

Commercial System

- Fast Data Finder [60]
  - Produced by Paracel, Inc.
- Text filtering based on user-specified profile
  - Special-purpose parallel hardware for speed
- Automatic word-by-word profile translation
  - Between English and Japanese
  - Users typically postedit the translations by hand

30

Cross-Language Information Retrieval

Corpus-based Techniques
for Free Text Searching

Types of Bilingual Corpora

- Parallel corpora: translation-equivalent pairs
  - Document pairs
  - Sentence pairs
  - Term pairs
- Comparable corpora
  - Content-equivalent document pairs
- Unaligned corpora
  - Content from the same domain

32

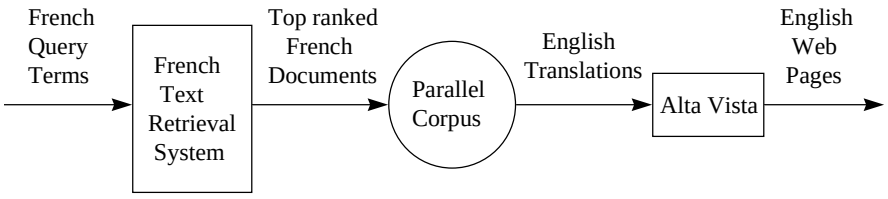
Generating Parallel Corpora

- Parallel corpora are naturally domain-tuned
  - Finding one for the right domain may be hard
- Alternative is to build one
  - Start with a monolingual corpus
  - Automatic machine translation for second language
- Worthwhile when IR technique is faster than MT
  - If translation errors don't hurt the IR technique
- Good results with Latent Semantic Indexing [27]

33

Pseudo-Relevance Feedback

- Enter query terms in French [31]
- Find top French documents in parallel corpus
- Construct a query from English translations
- Perform a monolingual free text search



34

Learning From Document Pairs

- Count how often each term occurs in each pair
  - Treat each pair as a single document

	English Terms					Spanish Terms			
	E1	E2	E3	E4	E5	S1	S2	S3	S4
Doc 1	4		2			2			1
Doc 2	8		4			4			2
Doc 3		2		2			2	1	
Doc 4		2	1				2		1
Doc 5	4				1	2		1	

35

Document and Term Similarity

- Doc 1 and Doc 2 use similar term patterns
  - After adjusting for the number of terms in each
  - This is how vector space retrieval works
    - Compute a term weight from each term count
    - Cosine: Normalize length, compute inner product
- Terms E1 and E3 are used in similar ways
  - Terms E1 & S1 (or E3 & S4) are more similar
    - Same computation reveals “term similarity”
    - Works both across and within languages

36

Similarity-Based Dictionaries

- Automatically developed from aligned documents
  - Reflects language use in a specific domain
- For each term, find most similar in other language
  - Retain only the top few (5 or so)
- Performs as well as dictionary-based techniques
  - Evaluated on a comparable corpus of news stories [98]
    - Stories were automatically linked based on date and subject

37

Generalized Vector Space Model

- "Term space" of each language is different
  - But the "document space" for a corpus is the same
- Describe new documents based on the corpus
  - Vector of cosine similarity to each corpus document
  - Easily generated from a vector of term weights
    - Multiply by the term-document matrix
- Compute cosine similarity in document space
- Excellent results when the domain is the same [11]

38

Latent Semantic Indexing

- Cosine similarity captures noise with content
  - Term choice variation and word sense ambiguity
- Signal processing to reduce term choice effect
  - Content-preserving dimensionality reduction
    - Conflate terms with similar usage patterns in corpus
      - Reduces several thousand dimensions to around 300
    - Linear mapping can be learned in O(corpus size) time
  - Applied to both documents and query

39

Latent Semantic Indexing

- Designed for better monolingual effectiveness
  - Works well across languages too [27]
    - Cross-language is just a type of term choice variation
- Produces short dense document vectors
  - Better than long sparse ones for adaptive filtering
    - Training data needs grow with dimensionality
  - Not as good for retrieval efficiency
    - Always 300 multiplications, even for short queries

40

Sentence-Aligned Parallel Corpora

- Easily constructed from aligned documents
  - Match pattern of relative sentence lengths
- Not yet used directly for effective retrieval [15]
  - But all experiments have included domain shift
- Good first step for term alignment
  - Sentences define a natural context

41

Cooccurrence-Based Dictionaries

- Align terms using cooccurrence statistics
  - How often do a term pair occur in sentence pairs?
    - Weighted by relative position in the sentences
  - Retain term pairs that occur unusually often
- Useful for query translation [11]
  - Excellent results when the domain is the same
- Also practical for document translation [67]
  - Term use variations to reinforce good translations

42

Exploiting Unaligned Corpora

- Documents about the same set of subjects
  - No known relationship between document pairs
  - Easily available in many applications
- Two approaches
  - Use a dictionary for rough translation
    - But refine it using the unaligned bilingual corpus
  - Use a dictionary to find alignments in the corpus
    - Then extract translation knowledge from the alignments

43

Feedback with Unaligned Corpora

- Pseudo-relevance feedback is fully automatic
  - Augment the query with top ranked documents
- Improves high-recall (find 'em all) effectiveness
  - "Recenters" queries based on the corpus
  - Short queries get the most dramatic improvement
- Two opportunities: [5]
  - Query language: Improve the query
  - Document language: Suppress translation error

44

Context Linking

- Automatically align portions of documents
  - For each query term:
    - Find translation pairs in corpus using dictionary
    - Select a "context" of nearby terms
      - e.g., +/- 5 words in each language
- Choose translations from most similar contexts
  - Based on cooccurrence with other translation pairs
- No reported experimental results [110]

45

Which to Use?

- Controlled vocabulary
  - Mature, efficient, easily explained
- Dictionary-based
  - Simple, broad coverage
- Comparable and parallel corpora
  - Effective in the same domain
- Unaligned corpora
  - Experimental

46

Cross-Language Information Retrieval

Performance Evaluation

Aspects of Performance

- Effectiveness
  - How well it finds what you asked for
- Efficiency
  - Retrieval: How it scales up to large collections
  - Filtering: How it handles high-volume streams
- Usability
  - How useful it is for finding what you want

Measuring Effectiveness

- Precision measures density
  - Fraction of documents in a set that are relevant
- Recall measures comprehensiveness
  - Fraction of relevant documents that are in the set
- Combine to produce a figure of merit
  - Traditional: Average precision at 11 recall values
  - Alta Vista: Precision at 20 documents
  - Known Item: Rank of the first relevant document

49

Sources of Measurement Error

- What does it mean to be relevant?
  - Relationship between an topic and a document
    - Identical copies of a relevant document are relevant
    - 80% inter-rater reliability is considered good
- Enormous variation across collections
  - Use the same test collection for each approach
    - Or report the cross-language to monolingual ratio
  - Average performance over lots of queries
    - Queries with proper names often do much better

50

Constructing Test Collections

- One collection for retrospective retrieval
  - Start with a monolingual test collection
    - Documents, queries, relevance judgments
  - Translate the queries by hand
- Need 2 collections for adaptive filtering
  - Monolingual test collection in one language
  - Plus a document collection in the other language
    - Generate relevance judgments for the same queries

51

Evaluating Corpus-Based Techniques

- Same domain evaluation [11]
  - Partition a bilingual corpus
  - Design queries
  - Generate relevance judgments for evaluation part
- Cross-domain evaluation [15]
  - Can use existing collections and corpora
  - No good metric for degree of domain shift

52

Evaluation Example

- Corpus-based same domain evaluation
- Use average precision as figure of merit

Technique	Cross-lang	Mono-lingual	Ratio
Cooccurrence-based dictionary	0.43	0.47	91%
Pseudo-relevance feedback	0.40	0.44	90%
Generalized vector space model	0.38	0.40	95%
Latent semantic indexing	0.31	0.37	84%
Dictionary-based translation	0.29	0.47	61%

From Carbonell, et al, "Translingual Information Retrieval: A Comparative Evaluation," IJCAI-97 [11]

53

TREC-6 Cross-Language Track

- Known item retrieval design
  - Queries are crafted to retrieve a known document
- 25 queries in three languages
  - English, French, German
- Several hundred MB of news in each language
  - Some comparable document alignments are known
- 11 groups planning to participate
  - Blind evaluation by NIST

54

Cross-Language Information Retrieval

System Integration

Integration Issues

- User Interface
  - Query formulation
  - Selection interface
  - Localization
- Document collection
  - Language identification
  - Content representation

Query Formulation

- Interactive word sense disambiguation [17]
- Show users the translated query
  - Retranslate it for monolingual users
- Provide an easy way of adjusting it
  - But don't require that users adjust or approve it

57

Selection Interface

- Document selection is a decision process [88]
  - Relevance feedback, problem refinement, read it
  - Based on factors not used by the retrieval system
- Provide information to support that decision
  - May not require very good translations
    - e.g., Word-by-word title translation
  - People can "read past" some ambiguity
    - May help to display a few alternative translations

58

Language Identification

- Can be specified using metadata
  - Included in HTTP and HTML
- Determined using word-scale features
  - Which dictionary gets the most hits?
- Determined using subword features
  - Letter n-grams in electronic and printed text
  - Phoneme n-grams in speech

59

Content Representation

- Electronic text
  - Character set (ASCII, Latin-1, Unicode, …)
- Document images
  - Skew removal
  - Font recognition
  - Word segmentation
- Speech
  - Recognition available for only a few languages

60

Cross-Language Information Retrieval

Research Directions

Research Directions

- User needs assessment
- Evaluation
- Corpus construction
- Word sense disambiguation
- System integration
- Probabilistic models
- Adaptive filtering

62

User Needs Assessment

- Who are the potential users? [105]
  - Some insight from controlled vocabulary systems
  - But free text on a worldwide network is different
- What goals do we seek to support? [74]
  - Known item, precision, recall, exploration, …
- What language skills must we accommodate?
  - Query formulation, browsing, comprehension

63

Evaluation

- Most critical need is for side by side tests
  - TREC-6 will do this using known item search
- Domain shift metric
  - Domain shift hurts corpus-based techniques
  - Need a way to measure severity of the shift
- Test collections for adaptive filtering
  - From cross-language recall/precision evaluation

64

Corpus Construction

- Corpus-based techniques have great potential
- Parallel corpora are rare and expensive
  - Find it, reverse engineer the links, clean it up
- Unlinked corpora are of limited value
  - Context linking research could change that [77]
- Comparable corpora offer middle ground
  - Need to develop automatic linking techniques
  - Also need a metric for degree of comparability

65

Word Sense Disambiguation

- Principal problem is translation ambiguity
- Some disambiguation strategies exist
  - Part of speech tagging
  - Phrase indexing
- Might benefit from a principled approach
  - Automatic disambiguation in documents
  - Selective interactive query disambiguation

66

System Integration

- User interface
  - Interactive refinement for query translation
    - Choosing which terms to seek help with
  - Codesign of search engine and user interface
    - Language skills may alter traditional division of work
  - Integration of on-demand translation
- Network interface
  - Language recognition for multilanguage documents

67

Probabilistic Models

- Vector space models have been popular
  - Easily constructed and visualized
  - Well suited to word-by-word translation
- Probabilistic models offer advantages
  - Sound way of exploiting corpus statistics
  - Richer basis for optimization
- Inquery is available for research use [5]

68

Adaptive Filtering

- Learn to predict user reaction to new documents
  - Based on prior response to documents in any language
- Differs from retrospective retrieval in 3 ways:
  - Unique effectiveness issues [73]
    - Uniform document representations for machine learning
  - Unique efficiency issues
    - Index profiles rather than documents
  - Unique usability issues [60]
    - User control over cross-language profile construction

69

Sponsored Research

- DARPA Information Technology Office
  - TIPSTER Phase III, BAA, SBIR
- NSF Interactive Systems Program
  - Stimulate, Digital Library Initiative
- European Community DG XIII
  - Language Engineering, Information Engineering, Telematics for Libraries

70

Cross-Language IR on the Web

- http://www.clis.umd.edu/dlrg/clir/
  - Most workshop proceedings
  - Lots of papers and project descriptions
  - Links to working systems
    - Including 2 web search engines
  - Useful linguistic resources
  - BibTeX for the attached bibliography

71

A Cross-Language IR Bookshelf

- Some useful things are not on the web
  - TIPSTER Phase II workshop proceedings [23]
    - And most other conference proceedings
  - Several important papers
    - CMU cross-system comparisons [5]
    - Everything from EMIR [85]
    - Controlled vocabulary work at Huddersfield [83]
    - Early Russian work [81]
  - ISO standard 5964 [46]

72

Where to From Here?

- International Joint Conference on AI (IJCAI)
  - August 23-29, 1997 in Nagoya, Japan
- European Conference on Digital Libraries
  - September 1-3, 1997 in Pisa, Italy
- Information Retrieval with Asian Languages
  - October 8-9, 1997 in Tsukuba-City, Japan
- Text Retrieval Conference (TREC-6)
  - November 19-21, 1997 in Gaithersburg, MD

73