

Cross Lingual Information Retrieval

Prasenjit Majumder

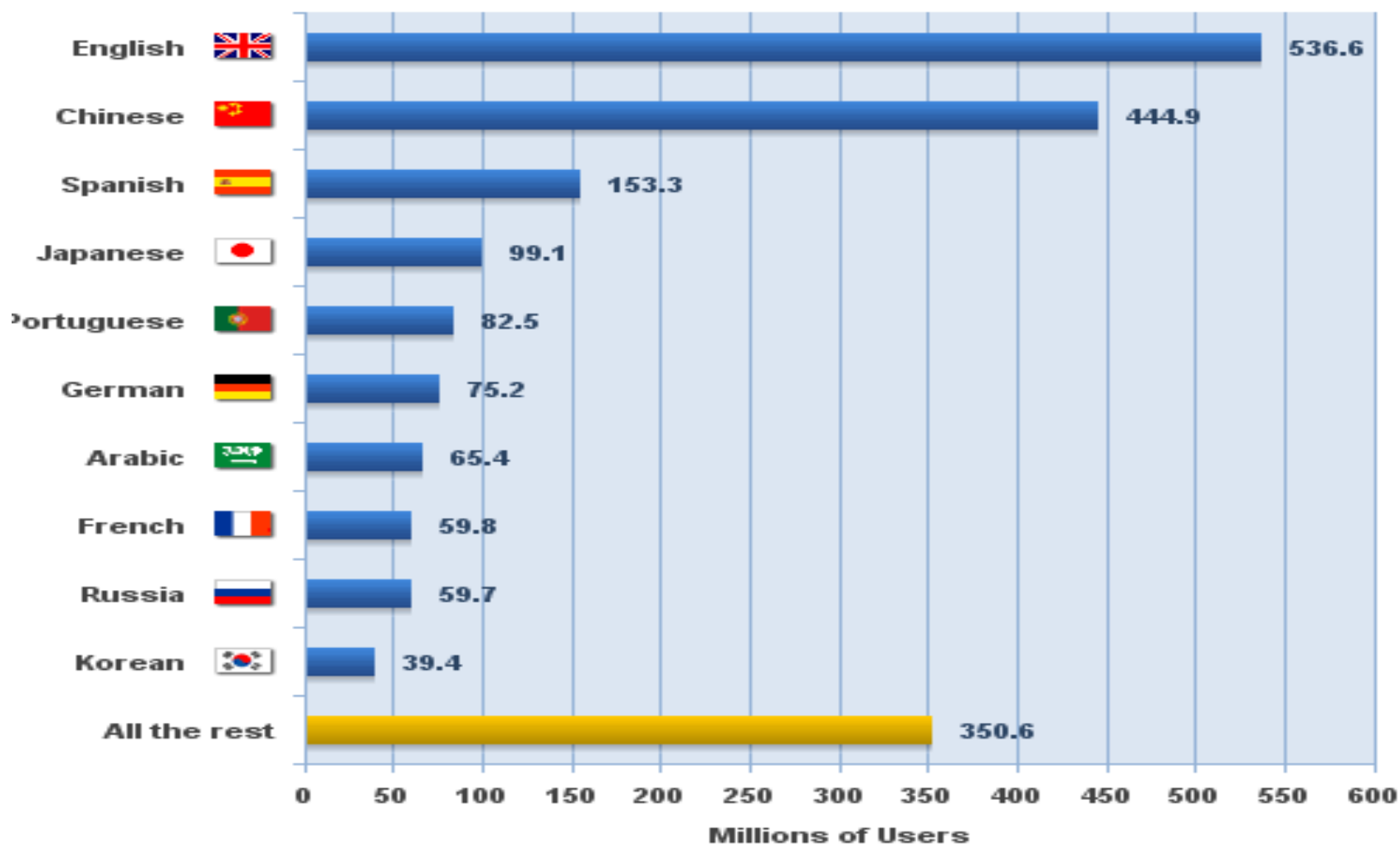
DAIICT, Gandhinagar

Yet Another Fire Fighter!

Talk Outline

- Why CLIR?
- What is CLIR?
- How to handle?
- Some interesting results

Top Ten Languages in the Internet 2010 - in millions of users



Source: Internet World Stats - www.internetworldstats.com/stats7.htm

Estimated Internet users are 1,966,514,816 on June 30, 2010

Copyright © 2000 - 2010, Miniwatts Marketing Group

The General Problem

Find documents written in any language

- Using queries expressed in a single language



يا ليلي يا عيني

Исследований



高等学校

att förstå

których można

The General Problem

- Traditional IR identifies relevant documents in the same language as the query (monolingual IR)
- Cross-language information retrieval (CLIR) tries to identify relevant documents in a language different from that of the query
- This problem is more and more acute for IR on the Web due to the fact that the Web is a truly multilingual environment

Cross-Language Text Retrieval

Query Translation

Document Translation

Text Translation

Vector Translation

Controlled Vocabulary

Free Text

Knowledge-based

Corpus-based

Ontology-based

Dictionary-based

Term-aligned

Sentence-aligned

Document-aligned

Unaligned

Thesaurus-based

Parallel

Comparable

Query vs. Document Translation

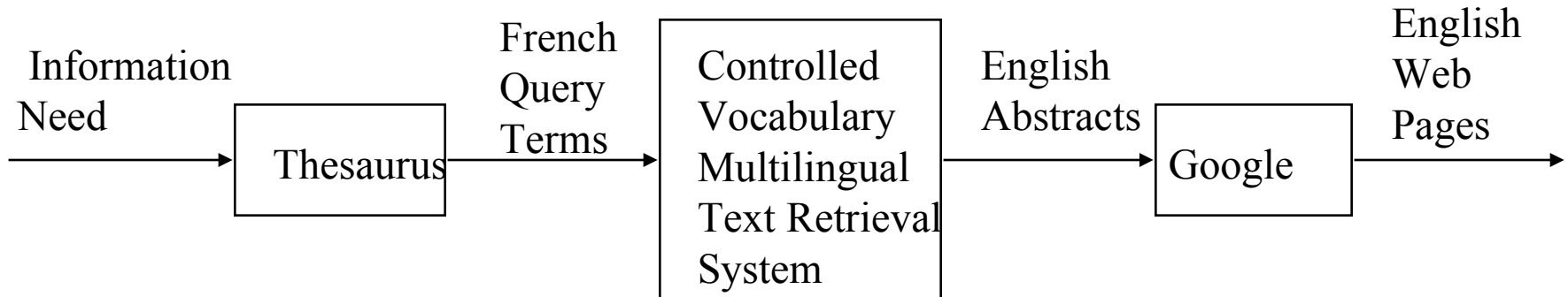
- Query translation
 - Very efficient for short queries
 - Not as big an advantage for relevance feedback
 - Hard to resolve ambiguous query terms
- Document translation
 - May be needed by the selection interface
 - And supports adaptive filtering well
 - Slow, but only need to do it once per document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

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

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

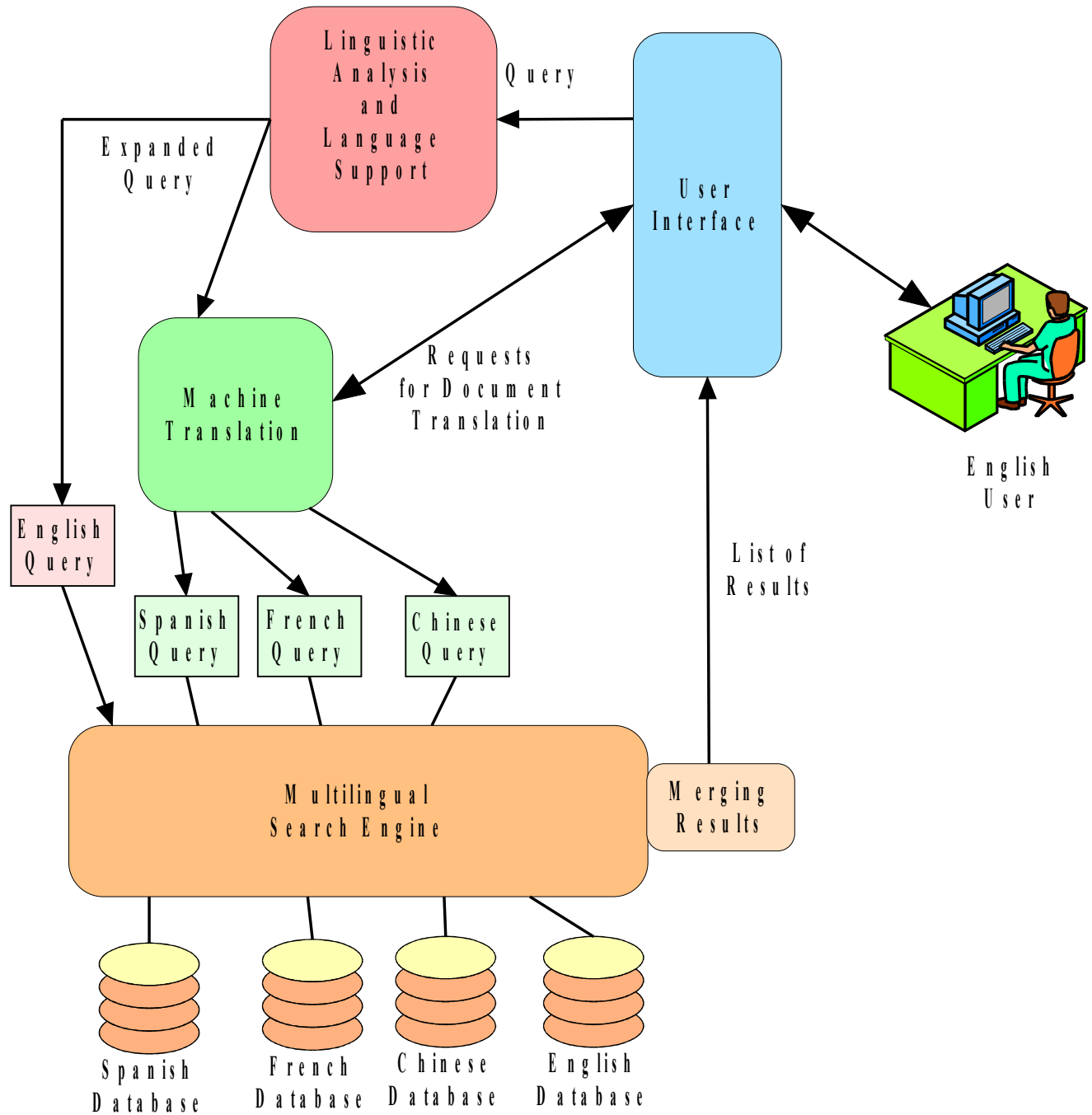


Machine Readable Dictionaries

- 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

Unconstrained Query Translation

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



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

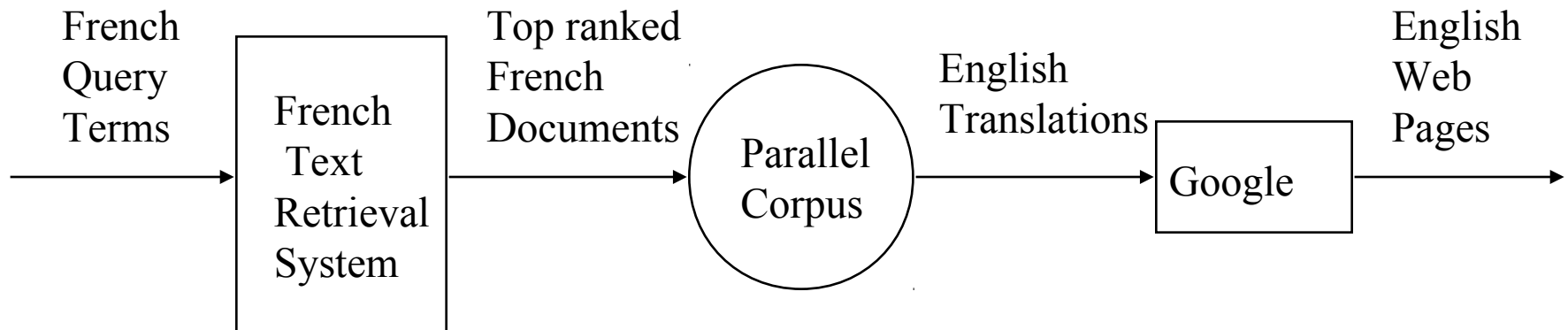
Corpus-based Techniques

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

Pseudo-Relevance Feedback

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



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	

Similarity-Based Dictionaries

- Automatically developed from aligned documents
 - Terms E1 and E3 are used in similar ways
 - Terms E1 & S1 (or E3 & S4) are even more similar
- 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
 - Stories were automatically linked based on date and subject

Sentence-Aligned Parallel Corpora

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

Cooccurrence-Based Translation

- 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
 - Excellent results when the domain is the same
- Also practical for document translation
 - Term usage reinforces good translations

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

Feedback with Unaligned Corpora

- Pseudo-relevance feedback is fully automatic
 - Augment the query with top ranked documents
- Improves recall
 - “Recenters” queries based on the corpus
 - Short queries get the most dramatic improvement
- Two opportunities:
 - Query language: Improve the query
 - Document language: Suppress translation error

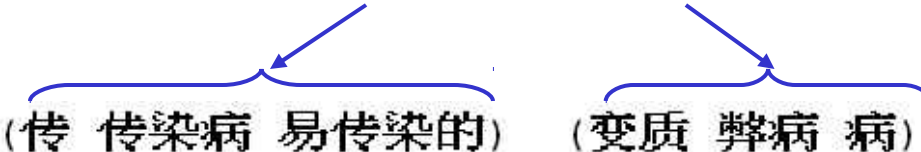
Sources of “Translation Knowledge”

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

Dictionary-Based Query Translation

Original query: El Nino and infectious diseases

Term selection: “El Nino” infectious diseases

Term translation:  (传 传染病 易传染的) (变质 弊病 病)

(Dictionary coverage: “El Nino” is not found)

Translation selection: (传染病 易传染的) 病

Query formulation:

Structure: OP1 (OP2 (传染病 易传染的) 病)

Computing Weights

- Unbalanced (#sum):
$$\frac{1}{3} \left[\frac{TF_1}{DF_1} + \frac{TF_2}{DF_2} + \frac{TF_3}{DF_3} \right]$$
 - Overweights query terms that have many translations
- Balanced (#sum(#sum)):
$$\frac{1}{2} \left[\frac{1}{2} \left(\frac{TF_1}{DF_1} + \frac{TF_2}{DF_2} \right) + \frac{TF_3}{DF_3} \right]$$
 - Sensitive to rare translations
- Structured (#syn):
$$\frac{1}{2} \left[\frac{TF_1 + TF_2}{DF_1 \cup DF_2} + \frac{TF_3}{DF_3} \right]$$
 - Deemphasizes query terms with any common translation

(Query Terms: 1: 传染病 2: 易传染的 3: 病)

“Structured Queries”

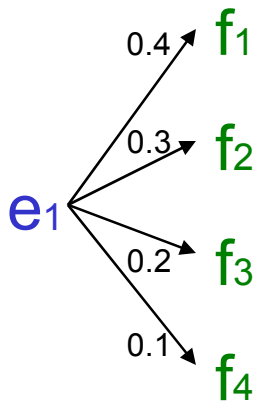
$$TF_j(q_i) = \sum_{f_k \in T(q_i)} TF_j(f_k)$$

$$DF(q_i) = \left| \bigcup_{f_k \in T(q_i)} \{D \mid f_k \in D\} \right|$$

Weighted Structured Queries (#wsyn)

$$tf(e_i, d_k) = \sum_{f_j} p(e_i \leftrightarrow f_j) * tf(f_j, d_k)$$

$$df(e_i) = \sum_{f_j} p(e_i \leftrightarrow f_j) * df(f_j)$$



f_j	f_1	f_2	f_3	f_4
$tf(f_j, d_k)$	20	5	2	50
$df(f_j)$	50	40	30	200
$p(e_i \leftrightarrow f_j)$	0.4	0.3	0.2	0.1
$tf(e_i, d_k)$	$0.4 * 20 + 0.3 * 5 + 0.2 * 2 + 0.1 * 50 = 14.9$			
$df(e_i)$	$0.4 * 50 + 0.3 * 40 + 0.2 * 30 + 0.1 * 200 = 58$			

Query Expansion in Cross-language IR

Ballesteros and Croft 1997

Use query expansion to improve results for Cross-lingual document retrieval

Translation is necessary but lowers performance:

- Machine Translation
- Parallel or Comparable Corpora techniques
- Machine readable dictionary (MRD)
 - Cheap and uncomplicated
 - Drops 40-60% below monolingual retrieval effectiveness

Causes for bad performance in MRD:

- Out of vocabulary words(e.g. technical terms)
- Addition of extraneous words to the translation
- Bad translation of multiterm phrases

Approaches using query expansion:

- Query expansion before translation
- Query expansion after translation
- Both before and after translation

Their experiment:

Comparison of retrieval using MRD without query expansion, with local feedback, and local context analysis query expansion.

Languages:

- Source Language: English
- Target Language: Spanish

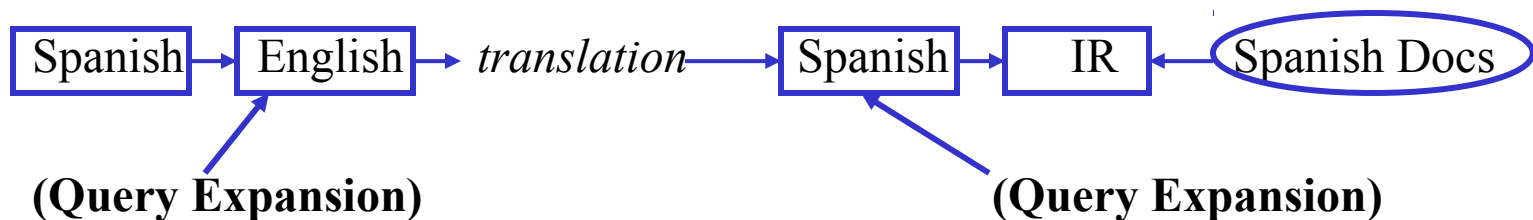
Collection:

- En Norte and San Jose Mercury News(208 and 301 Mb respectively)

IR System:

- INQUERY

Their system:



Pre-translation Query Expansion

(comparison of Local Feedback and Local Context Analysis)

- Collins Spanish-English MRD
- Phrase translation whenever possible. Otherwise word by word

Las relaciones economicas y comerciales entre Mexico y Canada

The economic and (commercial relations) between mexico and canada

Economic(commercial relations) mexico canada

Mexico(trade agreement)(trade zone)cuba salinas

[economico equitativo][comercio negocio trafico industria][narracion relato relacion][Mejico Mexico]Canada[Mejico Mexico]

1.- Original

2.- BASE (translation)

3.- LCA expanded BASE

4.- WBW + phrasal translation of LCA expanded BASE

Method	Avg.	% change
MRD	0.0826	
MRD + Phr	0.0826	0.3
MRD + LCA-WBW	0.0969	17.7
MRD + LCA-phr	0.1009	22.7
MRD + Phr+LCA-Phr	0.1053	27.9
LF	0.1099	33.5

--Phrase translation is beneficial

--Still LCA is less effective than LF(LCA more sensitive to wrong phrasal translations)

Post-translation Query Expansion

- Added concepts
Top-1 got a weight of 1.0
Each additional one's weight decreased by 1/100 of the previous one's weight
LF added top 20 concepts from the top 50 documents
LCA added top 100 concepts from the top 20 passages

Economic commercial relations mexico european countries

Comerc narr relat rel econom equit rentabl pai patri camp region tierr mej mex europ

(est un) canada pai europ francie (diversific comerc) mex polit pais alemani rentabl oportum
product apoy australi (merc eurp) agricultor bancarroto region (comun econom europ)

1.- Base translation

2.- MRD translation of BASE

3.- 20 post-translation LCA expansion

	MRD	LF	LCA 20
Avg prec	0.0824	0.0916	0.1022
% change		11.3	24.1

- Without bad phase translation factor, LCA seems to perform better than LF

Combined Pre- and Post-translation

- At Post-translation 50 top terms from 20 top passages were added

las relaciones economicas y comerciales entre Mexico y Canada

the economic and (commercial relations) between mexico and canada

economic(commercial relations) mexico canada mexico free-trade canada trade mexican
salinas Cuba pact economies barriers

[economico equitativo][comercio negocio trafico industria][narracion relato relacion]
[Mejico Mexico]Canada[Mejico Mexico][convenio comercial][comercio negocio trafico
industria]zona cuba salinas

canada(libr comerci) trat ottaw dosm (acuer paralel)norteamer(est un)(tres pais) import
eu (vit econom) comerci (centr econom)(barrer comerc)(increment subit)superpot rel
acuerd negoci

- 1.- Original
- 2.- BASE + Phr
- 3.- LCA expanded BASE
- 4.- WBD + phr translation
- 5.- LCA expanded translation

Results

	MRD	LF	LCA20-50
Avg prec	0.0823	0.1242	0.1358
% change		51.0	65.0

- Combined method is more effective than Pre- and Post-translation.
- LCA is better at precision (appropriate for Cross-Lingual IR)

Method	Precision	% Monolingual
Monolingual	0.1998	
MRD	0.0823	41.2
Pre-LF	0.1099	55.0
Pre-LCA	0.1139	57.0
Post-LF	0.0916	45.8
Post-LCA	0.1022	51.1
Comb-LF	0.1242	62.2
Comb-LCA	0.1358	68.0

- Machine Readable Dictionary translation to cheap, fast method for CLIR
- The quality of translation affects retrieval
- Poor phrasal translation decreases effectiveness in retrieval
- Both Local Feedback and Local Context Analysis reduce the effects of the poor translations
- LCA gives higher precision (particularly at low recall levels)
- Combined Pre- and Post- LCA expansion gives the best results
 - Reduces translation errors over 45% of MRD
 - From 42% to 68% of monolingual IR
 - Improvement in phrasal translation should help in reducing the gap

Bilingual X2EN: Indian Subtask

Track	Rank	Part.	Lang.	Experiment DOI	MAP
Hindi to English	1st	bombay-ltrc	hi	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	29.52%
	2nd	msindia	hi	10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.IITB_HINDI_TITLEDESC_DICE	21.80%
	3rd	hyderabad	hi	10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.IITB_HINDI_TITLEDESC_DICE	15.60%
	4th	jadavpur	hi	10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.IITB_HINDI_TITLEDESC_DICE	10.86%
	5th	kharagpur	hi	10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.IITB_HINDI_TITLEDESC_DICE	4.77%
	6th				
	Difference				
Bengali/Hindi/Marathi/Telugu to English	1st	bombay-ltrc	hi	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	29.52%
	2nd	msindia	hi	10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.IITB_HINDI_TITLEDESC_DICE	21.80%
	3rd	bombay-ltrc	mr	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.MARATHI_TITLEDESC_DICE	11.63%
	4th	hyderabad	te	10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.TELUGU_TITLEDESC_DICE	21.55%
	5th	jadavpur	te	10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.TELUGU_TITLEDESC_DICE	11.28%
	6th	kharagpur	bn	10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.BENGALI_TITLEDESC_DICE	7.25%
	Difference				

- limited linguistic resources
- phoneme-based transliterations to generate equivalent English queries
- stemmers and morphological analyzers if available

Giorgio M. Di numzio et al. CLEF 2007

Ref/Acknowledgements

- **QUERY AND DOCUMENT EXPANSION IN TEXT RETRIEVAL**, Clara Isabel Cabezas, University of Maryland College Park
- Internet World Stats, <http://www.internetworldstats.com/stats7.htm>
- Paul Clough, Bridging the language gap: making digital collections available to a multilingual society, presentation, 2005
- D.W. Oard, A Survey of Multilingual Text Retrieval. Computer Science Technical Report Series; Vol. CS-TR-3615. 1996
- Ari Pirkola, et al. Dictionary-Based Cross-Language Information Retrieval_ Problems, Methods, and Research Findings. Information Retrieval, Vol. 4. 2001
- Wikipedia. Related pages.
- Metamodel.com. What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model? <http://www.metamodel.com/article.php?story=20030115211223271>. 2004
- Miguel E. Ruiz, CLIR. Slides for school seminars. 2001
- Rada Mihalcea, Information Retrieval and Web Search. Class slides. 2007

•

Query?

- Translation
 - bilingual resources
 - Dictionary
 - Prob- Dictionary
 - MT systems (Babble fish, Google MT)
 - Corpus
- Transliteration
- Disambiguation
- Expansion
 - Pre
 - Post
 - Both