

Chapter 14 : Cross-Language Information Retrieval

Overview

- Definitions

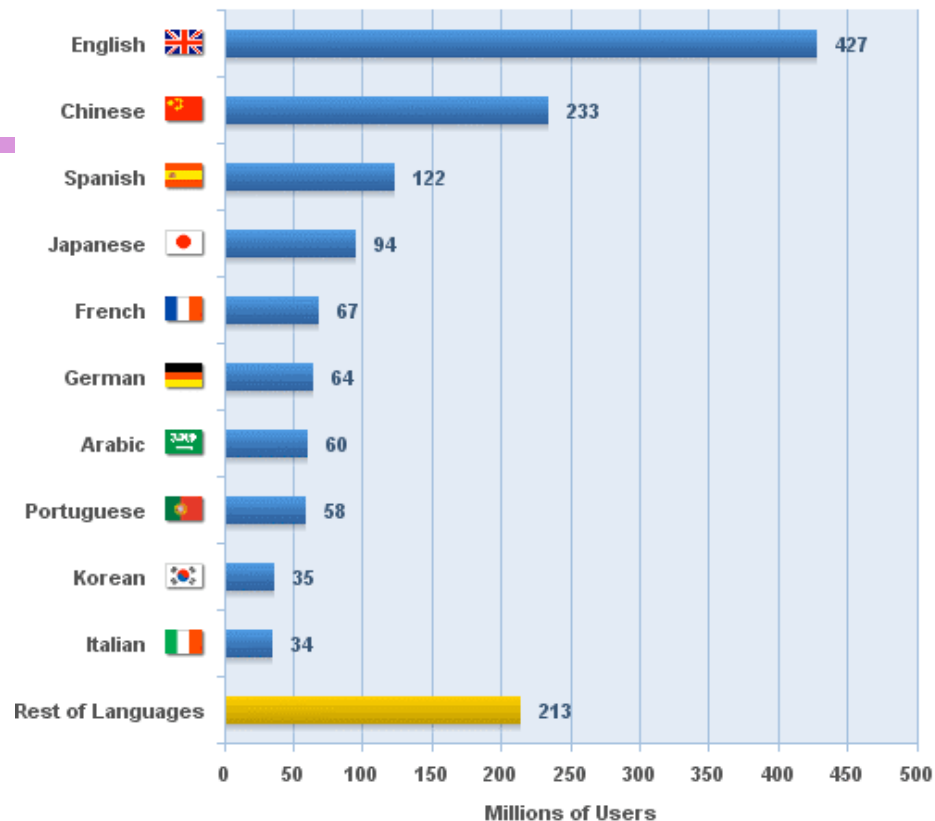
- Solutions CLIR:

- translation of the document collection in the language of the query
- **translation of the query in the language of the document collection:**
 - problem: word sense disambiguation and phrase translation
 - query translation approaches: use of
 - machine-readable dictionary (MRD)
 - machine translation (MT) system
 - parallel and comparable corpora
 - interlingua techniques

Definitions

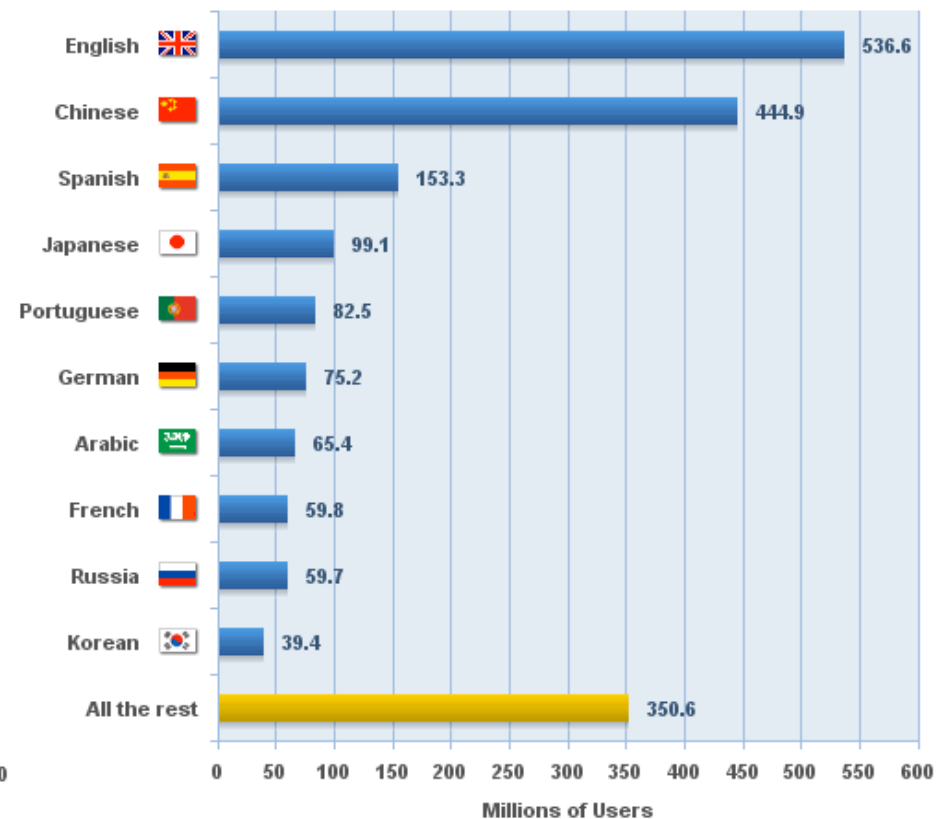
- Cross-language information retrieval = translingual information retrieval:
 - **concept**: query in one language (**source language**) and searching document collections in one or more different languages (**target languages**)
 - large attention due to increased accessibility of ever-more-diverse on-line international text collections (e.g., on the World Wide Web, digital libraries)
 - at the intersection of machine translation and information retrieval

Top 10 Internet Languages - May 2008 1,407,724,920 World Internet Users



Source: www.internetworldstats.com/stats7.htm
Copyright © 2008, Miniwatts Marketing Group

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

Translation of the document collection in the language of the query

- **Advantages:**

- better translation accuracy than when translating isolated queries
- better retrieval accuracy:
 - documents contain more information than queries: random translation errors should cause less degradation for the IR tasks in documents than in queries

- But: **practical feasibility is difficult:**

- large collections
- long computation and massive storage
- re-indexing of the translated collection

- If **summaries** are available: their translation could be a solution

Translation of the query in the language of the document collection

- Idea: retrieval of relevant documents based on the translated query, only the relevant documents might need to be translated afterwards
 - But: query translation = difficult:
 - queries are often short without internal structure (e.g., key terms)
 - often **different translations** are possible due to **word sense ambiguities**
 - **phrase translation** is hard
- = problems that are covered in this course unit

Other problems in CLIR

- Computational linguistic problems:
 - properly treating multiple character sets
 - language recognition
 - normalizing accentuation
 - language specific stemming routines or morphological analysis
 - ...
 - will not be covered

Word sense disambiguation

- Natural language words can have more than one semantic meaning or sense (polysemous and homonymous words)
- **Word sense disambiguation** can improve the precision and recall in information retrieval
- Word sense disambiguation includes the application of
 - knowledge of the syntactic class of the word (e.g., noun) obtained by part-of-speech tagging of the text
 - domain knowledge that relates a word class to a word meaning: considering the context in which the word occurs
= knowledge available in machine-readable dictionaries (MRDs)

Word sense disambiguation

- **Context** needed for disambiguation varies from
 - local context (e.g., words in the same sentence or surrounding sentences)
 - the complete text
 - the complete corpus (e.g., to disambiguate word senses in short texts of domain-specific corpus)

Word sense disambiguation

- Important clues:
 - **one sense per discourse**: the sense of a target word is highly consistent within a document
 - **one sense per collocation**: nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order and syntactic relationship

Word sense disambiguation in monolingual retrieval

Algorithm of Yarowski (1995): classical example of unsupervised word-sense disambiguation

w = an ambiguous word

$s_1, \dots, s_k, \dots, s_n$ = senses of the ambiguous word w

$c_1, \dots, c_j, \dots, c_m$ = contexts of w in a corpus

We use 2 sets for each sense s_k of the ambiguous word

w : F_k and E_k

F_k = **set of collocations** specific for sense s_k

collocation: here rule that describes a context (e.g. occurrence of word “live”)

each collocation has a score associated that corresponds to the likelihood that the collocation is specific to sense s_k

E_k = **the set of contexts** of the ambiguous word w with sense s_k

$C = \{c_1, \dots, c_m\}$

Algorithm:

1. Initialize C with the corpus: collect all the contexts of the ambiguous word in the corpus (= training examples)
2. For each sense s_k of w do

$F_k = \{ f_m \mid f_m \in \text{initial set of collocations} \}$

e.g., manually identifying a small number of seed collocations (“properly guessed”) representative for each sense

$E_k = \{ c_i \in C \mid \exists f_m : f_m \text{ applies to } c_i \wedge f_m \in F_k \}$

tagging all training examples containing the seed collocations with the seed's sense label

3. repeat until all E_k don't change anymore
for all senses s_k of w do

a) the classified training examples allow learning new collocations

$$F_k = \{ f_m \mid \forall n \neq k : \log \frac{P(s_k|f_m)}{P(s_n|f_m)} > \alpha \}$$

the strongest collocations for a particular context are chosen by selecting those collocations that are most characteristic of the just updated E_k

b) $E_k = \{c_i \in C \mid \exists f_m : f_m \text{ applies to } c_i \wedge f_m \in F_k\}$

the newly learned collocations allow tagging
(classifying) more training examples with the
sense label

c) optionally apply the constraint one sense per
discourse for augmenting and filtering the
classification

New text:

4) The learned dictionary (F_k) of collocations can be
applied to new data

Sense	Training examples
?	company said the <i>plant</i> is still operating
?	Although thousands of <i>plant</i> and animal species
?	zonal distribution of <i>plant</i> life
?	to strain microscopic <i>plant</i> life from the
?	vinyl chloride monomer <i>plant</i> , which is
?	and Golgi apparatus of <i>plant</i> and animal cells
?	computer disk drive <i>plant</i> located in
?	divide life into <i>plant</i> and animal kingdom
?	close-up studies of <i>plant</i> life and natural
?	Nissan car and truck <i>plant</i> in Japan is
?	keep a manufacturing <i>plant</i> profitable without
?	molecules found in <i>plant</i> and animal tissue
?	union responses to <i>plant</i> closures
?	animal rather than <i>plant</i> tissues can be
?	many dangers to <i>plant</i> and animal life
?	company manufacturing <i>plant</i> is Orlando
?	growth of aquatic <i>plant</i> life in water
?	automated manufacturing <i>plant</i> in Fremont
?	Animal and <i>plant</i> life are delicately
?	discovered at a St. Louis <i>plant</i> manufacturing
?	computer manufacturing <i>plant</i> and adjacent
?	the proliferation of <i>plant</i> and animal life
?	...

Figure 1: Step 1

[Yarowski ACL 1995]

Sense	Training Examples (Keyword in Context)
A	used to strain microscopic <i>plant</i> life from the ...
A	... zonal distribution of <i>plant</i> life
A	close-up studies of <i>plant</i> life and natural ...
A	too rapid growth of aquatic <i>plant</i> life in water ...
A	... the proliferation of <i>plant</i> and animal life ...
A	establishment phase of the <i>plant</i> virus life cycle ...
A	... that divide life into <i>plant</i> and animal kingdom
A	... many dangers to <i>plant</i> and animal life ...
A	mammals . Animal and <i>plant</i> life are delicately
A	beds too salty to support <i>plant</i> life . River ...
A	heavy seas, damage , and <i>plant</i> life growing on ...
A
?	... vinyl chloride monomer <i>plant</i> , which is ...
?	... molecules found in <i>plant</i> and animal tissue
?	... Nissan car and truck <i>plant</i> in Japan is ...
?	... and Golgi apparatus of <i>plant</i> and animal cells ...
?	... union responses to <i>plant</i> closures
?
?
?	... cell types found in the <i>plant</i> kingdom are ...
?	... company said the <i>plant</i> is still operating ...
?	... Although thousands of <i>plant</i> and animal species
?	... animal rather than <i>plant</i> tissues can be ...
?	... computer disk drive <i>plant</i> located in ...
B
B	automated manufacturing <i>plant</i> in Fremont ...
B	... vast manufacturing <i>plant</i> and distribution ...
B	chemical manufacturing <i>plant</i> , producing viscose
B	... keep a manufacturing <i>plant</i> profitable without
B	computer manufacturing <i>plant</i> and adjacent ...
B	discovered at a St. Louis <i>plant</i> manufacturing
B	... copper manufacturing <i>plant</i> found that they
B	copper wire manufacturing <i>plant</i> , for example ...
B	's cement manufacturing <i>plant</i> in Alpena ...
B	polystyrene manufacturing <i>plant</i> at its Dow ...
B	company manufacturing <i>plant</i> is in Orlando ...

[Yarowski ACL 1995]

Initial decision list for <i>plant</i> (abbreviated)		
LogL	Collocation	Sense
8.10	<i>plant</i> life	⇒ A
7.58	manufacturing <i>plant</i>	⇒ B
7.39	life (within ± 2 -10 words)	⇒ A
7.20	manufacturing (in ± 2 -10 words)	⇒ B
6.27	animal (within ± 2 -10 words)	⇒ A
4.70	equipment (within ± 2 -10 words)	⇒ B
4.39	employee (within ± 2 -10 words)	⇒ B
4.30	assembly <i>plant</i>	⇒ B
4.10	<i>plant</i> closure	⇒ B
3.52	<i>plant</i> species	⇒ A
3.48	automate (within ± 2 -10 words)	⇒ B
3.45	microscopic <i>plant</i>	⇒ A
	...	

[Yarowski ACL 1995]


Change in tag	Disc. no	Training examples
A->A	724	...the existence of plant and animal life...
A->A	724	...classified as either plant or animal...
?->A	724	Although bacterial and plant cells are enclosed
A->A	348	...the life of the plant, producing stem...
A->A	348	...an aspect of plant life, for example...
?->A	348	...tissues; because plant egg cells have...
?->A	348	photosynthesis, and so plant growth is attuned

Figure 1: Steps 3b and c Labeling previously untagged contexts and applying the one-sense-per-discourse constraint

Change in tag	Disc. no	Training examples
A->A	525	contains a varied plant and animal life
A->A	525	the most common plant life, the
A->A	525	slight within Arctic plant species
B->A	525	are protected by plant parts remaining from

Figure 2: Step 3c Error correction

[Yarowski ACL 1995]



LogL	Collocation	Sense
10.12	plant growth	=>A
9.68	car (within $\pm k$ words)	=>B
9.64	plant height	=>A
9.61	union (within $\pm k$ words)	=>B
9.54	equipment (within $\pm k$ words)	=>B
9.51	assembly plant	=>B
9.50	nuclear plant	=>B
9.31	flower (within $\pm k$ words)	=>A
9.24	job (within $\pm k$ words)	=>B
9.03	fruit (within $\pm k$ words)	=>A
9.02	plant species	=>A

Figure 1: Step 4 - final decision list for plant

[Yarowski ACL 1995]

Word sense disambiguation in CLIR

- Different approaches (see below) that take into account:
 - 1) Use of part-of-speech (POS) tags
 - 2) Use of co-occurrence statistics in corpus
 - 3) Use of query expansion
 - 4) Use of parallel corpus

Query translation approaches

- Use of
 - machine-readable dictionary (MRD)
 - machine translation (MT) system
 - parallel corpora

Translation of query with MRD

- Concept:
 - query is often formulated with single words or short phrases
 - each query word is translated by a bilingual dictionary
- Advantage:
 - dictionaries are often available
- Disadvantage:
 - experiments show a drop of about 50% in average precision when compared to mono-lingual performance

Translation of query with MRD

- Problems:

- 1) **Inadequate coverage of the dictionary:**

- missing word forms
- missing spelling variants, especially of proper names: e.g., “Yeltsin” (English), “Elstine” (French)

- 2) **Word sense disambiguation:** difficult by lack of

- knowledge of syntactic class of query term: queries are often not well-formed sentences
- context

Translation of query with MRD

3) translation of phrases

- difficult identification: query is often not a well-formed sentence
- translation:
 - word-by-word translation often results in incompatible meaning
 - not always one-to-one relation between term and its translation word (e.g., “enter into a discussion with” (English) and “entamer” (French))
 - dictionary with phrase translations = rare

Translation of query with MRD

- Solutions:

1. **Selection of the correct sense** for the translation of each query term

- computation of term by term association matrix based on reference collection in target language (e.g. term correlations based on Dice similarity, pointwise mutual information statistic)

- **word sense disambiguation:**

- given the set of original query terms
- select for each of the terms the best sense, such that the resulting set of selected senses contains senses that are most mutually related

■ Example:

Term association matrix of French document collection:

	budget	boucle	cabinet	état	meuble
budget		0.2	0.7	0.7	0.4
boucle			0.1	0.2	0.6
cabinet				0.8	0.7
état					0.6
meuble					

....

Query: “Budget of the cabinet”

Senses of the dictionary and translation of “cabinet” : meuble,
cabinet

Possible translations and sum of pair wise term correlation:

“Budget du cabinet”: 0.7

“Budget du meuble”: 0.4

Query: “Budget of the cabinet of the state”

Possible translations and sum of pair wise term correlations:

“Budget du cabinet de l'état”: $0.7 + 0.7 + 0.8 = \underline{2.2}$

“Budget du meuble de l'état”: $0.4 + 0.6 + 0.7 = 1.7$

-
- The above example shows that given the measurements of term similarity, ideally we should select for each query term the translation that co-occurs the most often with (or the most similar to) the selected translations of other terms in the same query
 - Finding such an optimal translation can become computationally very expensive, instead one uses an approximate greedy algorithm as follows

-
- Given a query in a source language (e.g., English) $Q = \{e_1, e_2, \dots, e_n\}$
 - For each query term e_i we define a set of m distinct translations (e.g., in Chinese) C : $C(e_i) = \{c_{i1}, c_{i2}, \dots, c_{im}\}$ according to a bilingual dictionary
 - For each set $C(e_i)$
 - For each translation $c_{ij} \in C(e_i)$, define the similarity score between the translation c_{ij} and a set $C(e_k)$ ($k \neq i$) set as the sum of the similarities between c_{ij} and each translation in the set $C(e_k)$ according to:

$$sim(c_{ij}, C(e_k)) = \sum_{c_{kl} \in C(e_k)} sim(c_{ij}, c_{kl})$$

-
- Compute the cohesion score for c_{ij} as

$$cohesion(c_{ij}|e, C) = \log \left[\sum_{C(e_k)} sim(c_{ij}, C(e_k)) \right]$$

- Select the translation $c \in C(e_i)$ with the highest cohesion score

$$c = \operatorname{argmax}_{c_{ij} \in C(e_i)} cohesion(c_{ij}|e, C)$$

- Greedy search: the translation of different query terms is independently determined
- Widely used algorithm

Translation of query with MRD

2) **Query expansion** with **all the possible translations**:

■ example:

- source language query: “Waldheim affair”
- translated into French:
 - “Waldheim adventure business affaire case liaison”

■ problem: query fed into the retrieval engine

- documents mentioning some of the last five words of translated query might be ranked higher than those mentioning only “Waldheim”
- additional mechanisms needed

Translation of query with MRD

2a) Translation into Boolean (structured) query :

- Boolean disjunction (\vee) is a natural way to link together many translation equivalents
- Boolean conjunction (\wedge) is likely to be an effective strategy for disambiguation:
 - assumption: correct translation equivalents of two or more query terms are much more likely to co-occur in the target language documents than any other corresponding incorrect translation equivalents
 - example: Waldheim \wedge (adventure \vee business affaire \vee case \vee liaison)

Translation of query with MRD

2b) Query expansion with pseudo-relevance feedback

■ **Pre-translation** feedback:

- supposes training corpus in language of query for initial retrieval
- expansion of the query with terms from the top ranked documents
- strengthens the initial query
- improves precision of the retrieval
- but, can introduce extraneous terms

■ **Post-translation** feedback:

- retrieval of documents from the target collection

Translation of query with MRD

- expansion of the query with terms of the top ranked documents from the collection after translation
- even if there is sufficient context, some terms are translated incorrectly
- decreases ambiguity by de-emphasizing inappropriate terms
- tends to improve recall
- **Combined** feedback
 - significant improvement especially for short queries: results close to monolingual retrieval
 - but, ambiguity arising from the word-by-word translation of phrases is still a major factor in the loss of performance

Language model

- **Language model for monolingual retrieval:**

$$P(Q_1, \dots, Q_m | D) = \prod_{i=1}^m ((1 - \lambda)P(Q_i | EC) + \lambda(P(Q_i | D)))$$

- **Language model for cross-lingual retrieval:**

$$P(E_1, \dots, E_m | D) = \prod_{i=1}^m ((1 - \lambda)P(E_i | EC) + \lambda \sum_{k=1}^t P(C_k | D)P(E_i | C_k))$$

where E_i = query term in the source language (e.g., English)

EC = corpus in the source language

C_k = is a translation of query term E_i (e.g., Chinese)

t = number of possible translations for E_i

Language model

- Language model takes into account **probabilistic term translation** learned from MRD or parallel corpus:
 - $P(E_i | C_k)$ = the probability of translation into the English word E_i given a Chinese word C_k
 - e.g., estimated from MRD: if the Chinese word $C_k = c_k$ has l translations e_1, \dots, e_l : $P(E_i = e_j | C_k = c_k) = 1/l$
 - e.g., estimated from aligned parallel corpus: cf. [Brown et al. CL 1993]

Language model and relevance feedback

- A cross language relevance model R is built as:

$$P(C \mid R_c) = \sum_{D_c \in R_c} P(D_c) P(C \mid D_c) \prod_{i=1}^m ((1 - \lambda) P(E_i \mid EC) + \lambda \sum_{k=1}^t P(C_k \mid D_c) P(E_i \mid C_k))$$

where

C = word in target language (e.g., Chinese word) (1)

R_c = set of relevant documents (e.g., in Chinese)

Language model and relevance feedback

$P(C=c|D_c)$ is estimated as:

$$\alpha \left(\frac{tf_{c,D_c}}{\sum_{w \in D_c} tf_{w,D_c}} \right) + (1 - \alpha) P(C=c|CC) \quad (2)$$

where

tf = term frequency

CC = corpus in target language (e.g., Chinese corpus)

w = word in target language

Language model and relevance feedback

- Once the cross-language relevance model (R) is estimated (e.g., Chinese relevance model for an English query):
 - documents of the collection are reranked: e.g., by increasing cross-entropy of R and D
 - the cross-entropy measures divergence between two language models:

$$H(p, q) = - \sum_x p(x) \log q(x)$$

- $CrossEntropy(R, D) = - \sum_w P(C = w | R) \log P(C = w | D)$

where $P(C=w|R)$ is estimated according to (1)

$P(C=w|D)$ is estimated according to (2)

Language model and relevance feedback

- In other words, we learn a language model for the query:
 - In the above example: a language model for the Chinese query, based on the relevant Chinese documents
 - Other ranking functions that compare probability distributions can be used (e.g., Kullback-Leibler divergence - see exercises)
 - The method can also be used in monolingual retrieval with relevance feedback

Translation of query with MT system

- **When a query is in form of textual description: use of a machine translation (MT) system:**
 - computerized system responsible for the production of translations from one natural language into another, with or without human assistance
 - natural language processing of lexical, syntactical and semantic properties of the text
 - perfection in automatic text understanding and translation is currently not realized
- **Problems of query translation:**
 - when queries are short or ill-formed: use of MRD is better suited

Learning from parallel corpora

■ **Parallel corpora:**

- two or more corpora with the same text written in different languages
- are usually aligned on a sentence level, but can be aligned on a word or phrase level
- in the techniques below: use of a bilingual corpus in source and target language

Text alignment

- = identifying which text strings in one language correspond to which text strings in parallel text of other language by being the translation of each other
- Alignment of:
 - sentences and paragraphs
 - words and phrases: more difficult
- Use of statistical techniques (not covered in this course)

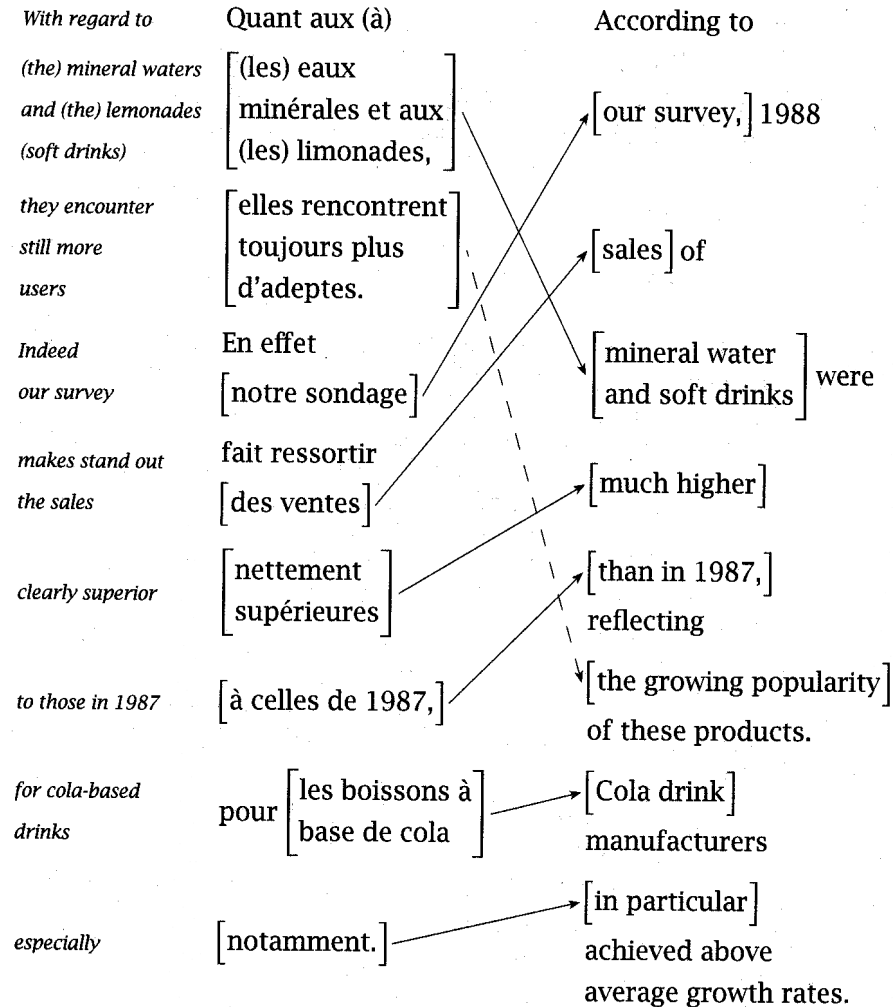


Figure 13.2 Alignment and correspondence. The middle and right columns show the French and English versions with arrows connecting parts that can be viewed as translations of each other. The italicized text in the left column is a fairly literal translation of the French text.

[Manning & Schütze 1999]

Learning from parallel corpora

- Query translation with parallel corpora:
 - word and phrase alignment:
 - translation of phrases
 - translation of words in context
- Query translation and expansion with parallel corpora
 - local strategy
 - global strategy
- Transformation of query and document representations based on parallel corpora: e.g.,
 - LSI

Query translation and expansion with parallel corpora

- Local strategy: **pseudo-relevance feedback using parallel corpora**
 - initial retrieval from source language corpus:
 - query in source language
 - selection of the top ranking documents
 - substitution of the documents with their counterpart documents in the target language
 - translated and expanded query = top-ranking documents in target language
- => used to query other corpus in target language
= finding similar documents
(= technique called PRF in graph below)

Query translation and expansion with parallel corpora

- Global strategy: **example based machine translation**:
 - uses a large corpus of example pairs of previously translated sentences, in order to find close matches and translations of words and phrases in context
- construction of **dictionary of translation candidates**:

construction of association matrix $A_{n \times m}$ where n = number of terms in source language and m = number of terms in target language

cell in A = number of times the source language word occurs in the same sentence pair as the target language word

Query translation and expansion with parallel corpora

if correlation value $>$ threshold: matching pair (x_i, y_i) is added to the dictionary where x_i is term in source language and y_i is term in target language

- **translation of the query:**

- composing words are sought in this dictionary
- all candidate translations are combined
(technique called EBT in graph below)

- technique can be improved with word and phrase alignment

Interlingua techniques

- Query and documents are converted to intermediate representation: e.g.,
 - Latent Semantic Indexing trained on parallel corpus
 - use of WordNet synset numbers (not covered in this chapter)

Transformation of query and document representations based on parallel corpora

- **Latent semantic indexing (LSI):** $\begin{bmatrix} A \\ B \end{bmatrix}$
- use of a **term-document matrix**:
 - row: each word from both languages in a parallel corpus
 - column: corresponding to document numbers
 - cell: corresponds to the number of times the word appears in a document
 - documents which are translations of each other have the same number(= cf. considering text and its translation as 1 document)

	•	doc1	doc2	doc3	doc4	doc5	doc6	...
• horse		1	0	2	0	0	0	...
• cheval		1	0	2	0	0	0	
• cabinet		1	2	0	0	0	0	
• meuble		1	0	0	0	0	0	
•						

Transformation of query and document representations based on parallel corpora

- **LSI :**

- matrix decomposition technique: singular-value decomposition (SVD):

$$\begin{bmatrix} A \\ B \end{bmatrix} = U_2 \Sigma_2 V_2^t$$

$$L_2 = U_2 \Sigma_2^{-1}$$

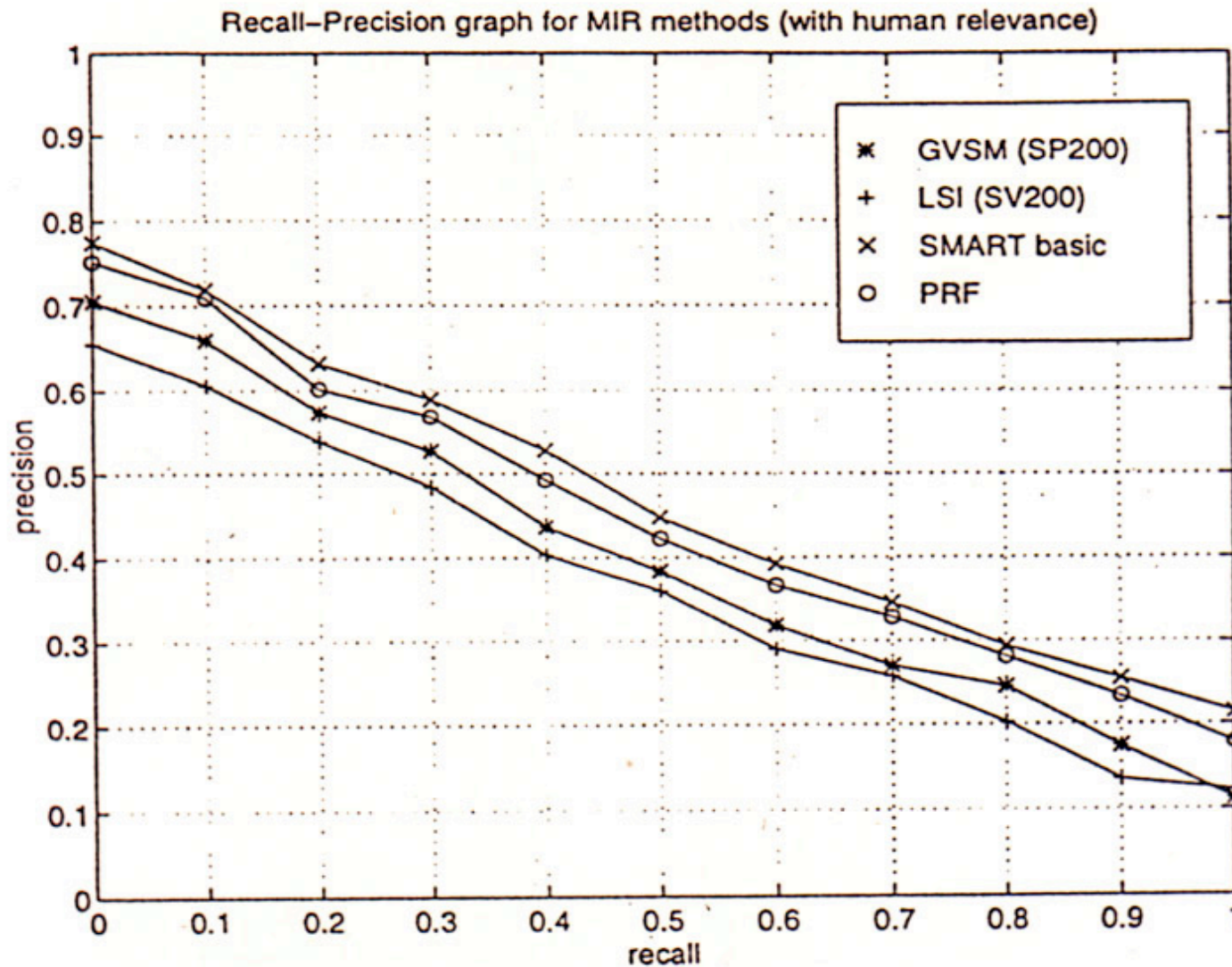
- where U_2 , V_2 and Σ_2 are the matrices computed using the singular value decomposition of the bilingual input matrix
- L_2 : represented in reduced (latent) semantic space

Transformation of query and document representations based on parallel corpora

The similarity of the query \mathbf{q} to the document \mathbf{d} in the latent semantic space is computed e.g., as:

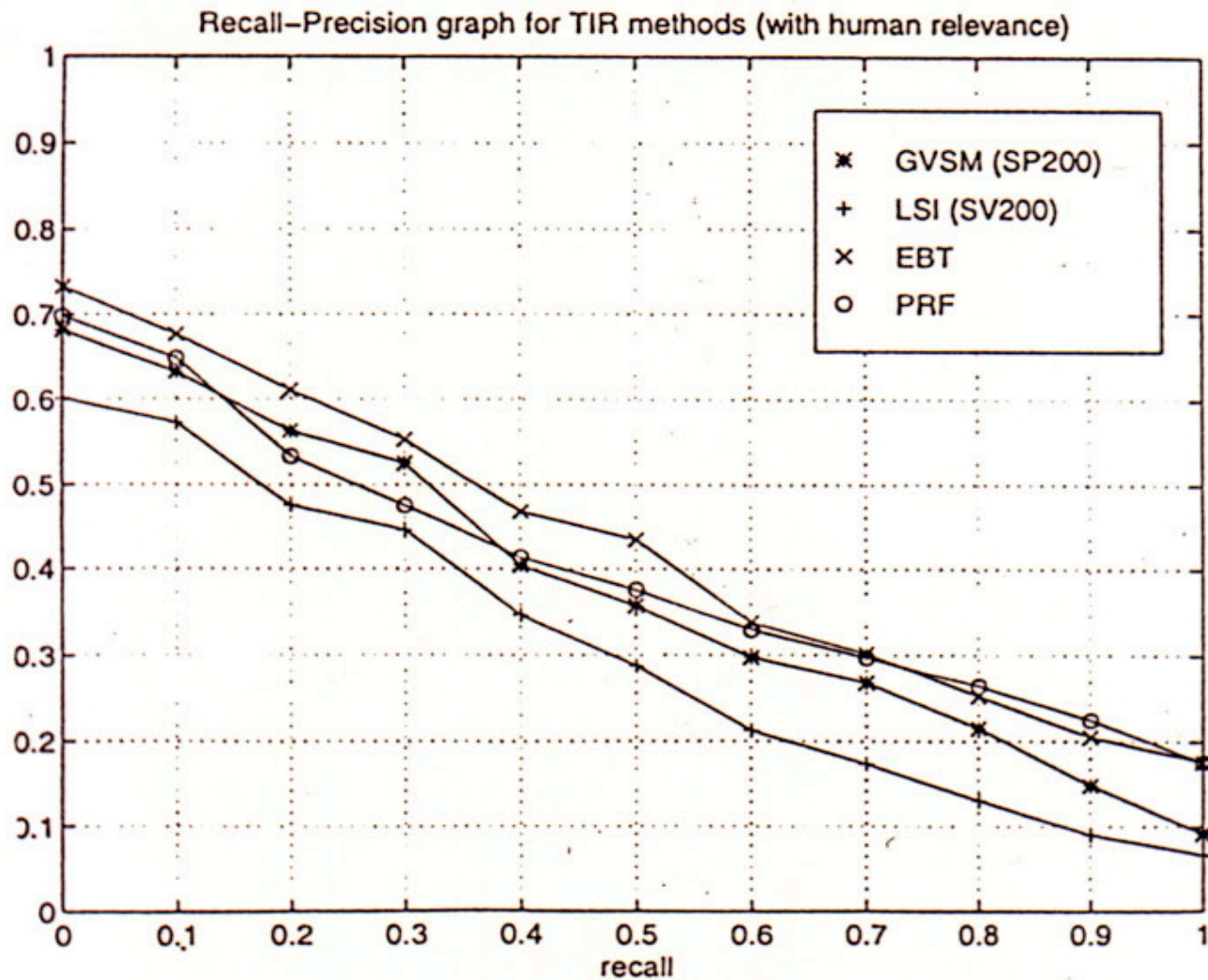
$$\text{sim}(\mathbf{q}, \mathbf{d}_j) = \cos(L_2^T \mathbf{q}, L_2^T \mathbf{d}_j)$$

- query and documents are expanded with their “synonyms” (here also their translations)
- retrieval:
 - query: in either language
 - vector comparison with language-independent representation of documents
 - retrieval of most similar documents regardless of the language



[Carbonell et al. IJCAI 1997]

Recall-Precision performance of MIR methods



[Carbonell et al. IJCAI 1997]

Recall-Precision performance of TIR methods

Learning from parallel corpora

- Advantages
 - generally improves CLIR:
 - experiments with Spanish and English UNICEF reports:
 - best results: EBT
 - disappointing LSI: supposes that a word only has one global sense in the parallel corpus and that the sense's translations in different languages map into the same point in thematic space
- Disadvantages:
 - restricted availability
 - results dependent on how well the corpora are aligned
 - no guarantee that all concepts of the query are covered

Learning from comparable corpora

- Also techniques proposed that learn from comparable corpora
 - **comparable corpora:**
 - two or more corpora that treat similar content written in different languages
 - often in restricted subject domain and written in “sublanguage”: language with restricted lexical, syntactical, semantic and discourse properties
 - results less encouraging

More CLIR results

Monolingual	Query translation	Document translation	Query and document translation	Probabilistic query translation
0.4275	0.2943	0.3197	0.3266	0.3615

Table 6.5. Comparing MT (SYSTRAN) and probabilistic translation using a language modeling approach (results are in terms of mean average precision (MAP)): source language is English, target language is Chinese.

[Xu & Weischedel 2003]

Table 3.

English–Arabic CLIR. Mean average precision on 25 TREC 2001 queries

	Dictionary	Unexpanded		Expanded		
		INQ	LM	INQ	LM	Rel
Monolingual		0.4135	0.3792	0.4367	0.4174	0.4109
Cross-lingual	UMass (nonparallel)	0.3807	0.3132	0.4518	0.4047	0.3897
	Parallel corpus	0.3160	0.3057	0.3812	0.4131	0.3958
	Combined	0.3809	0.3488	0.4443	0.4540	0.4432
% of Mono	Combined	92	92	102	109	108

Table 4.

[Larkey & Connell IP & M 2005]

English–Arabic CLIR. Mean average precision on 50 TREC 2002 queries

LM = language model

Rel = LM with relevance model: new query is made with 500 terms from top 20 (monolingual) and top 50 (cross-lingual) documents

Table 5.

English–Spanish CLIR. Mean average precision on 25 TREC 4 queries

	Dictionary	Unexpanded		Expanded		
		INQ	LM	INQ	LM	Rel
Monolingual		0.4994	0.4838	0.5259	0.5188	0.4845
Cross-lingual	Collins (nonparallel)	0.3596	0.3250	0.3900	0.4159	0.4407
	Parallel corpus	0.4144	0.4023	0.4775	0.4690	0.4830
	Combined	0.4024	0.3972	0.4464	0.4681	0.4813
% of Mono	Parallel	83	83	91	90	100

[Larkey & Connell IP & M 2005]

What have we learned?

- Most important model in CLIR: translation of the query
- Queries are short and are not often well-formed sentences:
 - difficult to translate correctly with translation dictionaries or machine translation systems
 - most important problems:
 - missing coverage of the dictionaries
 - disambiguation of word senses
 - translation of phrases
- Query expansion with terms learned from reference corpus or parallel corpora: decreases the ambiguity of translation
- When available, learning from parallel corpora:
 - eliminates the need for hand-built translation dictionaries: phrase dictionaries seldom exist anyway

Research questions to be solved

- Word and phrase alignments in parallel corpora of limited size
- Word and phrase alignments in comparable bilingual corpora
- Efficient computations of translation of query composed of several words given based on collocation patterns in target corpus

■ **International competitions:**

- TREC: English, Spanish, Chinese, German, French, Italian and Arabic
- CLEF: English, French, German, Italian, Swedish, Spanish, Dutch, Finnish and Russian
- NTCIR: English, Japanese, Chinese and Korean

Further reading

- Brown, P.E., Della Pietra, V.J., Della Pietra, S.A. & Mercer, R.L. (1993). The mathematics of statistical machine translation. *Computational Linguistics*, 19 (2), 263-311.
- Carbonell, J.G., Yang, Y., Frederking, R.E., Brown, R.D., Geng, Y. & Lee, D. (1997). Translingual information retrieval: A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence IJCAI-97* (pp. 708-714). San Francisco, CA: Morgan Kaufmann.
- Gao, J. & J.-Y. Nie (2006). A study of statistical models for query translation: Finding a good unit of translation. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 194-201). New York: ACM.
- Grefenstette, G. (1998). *Cross-Language Information Retrieval* (The Kluwer International Series on Information Retrieval) (pp. 1-9). Boston: Kluwer Academic Publishers.
- Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. *Information Processing & Management*, 41 (3), 433-455.
- Larkey, L.S. & Connell, M.E. (2005). Structured queries, language modeling, and relevance modeling in cross-language information retrieval. *Information Processing & Management*, 41(3), 457-473.

Lavrenko, V., Choquette, M. & Croft, W.B. (2002). Cross-lingual relevance models. In *Proceedings of the Twenty-Fifth annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 175-182). New York: ACM.

Manning, C.D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Boston, Ma: MIT Press. (pp. 244 -252)

McNamee, P. & Mayfield, J. (2002). Comparing cross-language expansion techniques by degrading translation resources. In *Proceedings of the Twenty-Fifth Annual International ACM SIGIR Conference on Information Retrieval* (pp. 159-166).

Oard, D.W. & Diekema, A.R. (1998). Cross-language information retrieval. *Annual Review of Information Science and Technology*, 33, 223-256.

Xu, J. & Weischedel, R. (2003). A probabilistic approach to term translation for cross-language information retrieval. In W.B. Croft & J. Lafferty (Eds.), *Language Modeling for Information Retrieval* (pp. 125-140). Boston, MA: Kluwer Academic Publishers.

Yarowski, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods.

[http:// citeseer.nj.nec.com/yarowski95unsupervised.html](http://citeseer.nj.nec.com/yarowski95unsupervised.html)

-
- How to find parallel corpora on the Web:
 - Resnik, P. (1999). Mining the Web for bilingual texts. In *Proceedings of ACL* (pp. 527-526).