CROSS-LANGUAGE INFORMATION RETRIEVAL AND BEYOND

Jian-Yun Nie
University of Montreal
http://www.iro.umontreal.ca/~nie

Outline

- What are the problems in CLIR?
- Recall: General approaches to IR
- The approaches to CLIR proposed in the literature
- Their effectiveness
- Remaining problems
- Applications

Problem of CLIR

- Cross-language IR (CLIR)
  - Use a query in a language (e.g. English) to retrieve documents in another language (Chinese)
- Multilingual IR (MLIR)
  - Use a query in one language to retrieve documents in several languages

History

- In 1970s, first papers on CLIR

• TREC-3 (1994)	Spanish (monolingual): *El Norte* Newspaper	SP 1-25
• TREC-4 (1995)	Spanish (monolingual): *El Norte* Newspaper	SP 26-50
• TREC-5 (1996)	Spanish (monolingual): *El Norte* newspaper and *Agence France Presse*	SP 51-75
	Chinese (monolingual): *Xinhua News agency, People's Daily*	CH 1-28
• TREC-6 (1997)	Chinese (monolingual), The same documents as TREC-6	CH 29-54
	CLIR:	
	English: *Associated Press*	CL 1-25
	French, German: *Schweizerische Depeschenagentur (SDA)*	
• TREC-7 (1998)	**CLIR**:	
	English, French, German, Italian *(SDA)*	CL 26-53
	+ German: *New Zurich Newspaper (NZZ)*	
• TREC-8 (1999)	**CLIR** (English, French, German, Italian): as inTREC-7	CL 54-81
• TREC-9 (2000)	English-Chinese: Chinese newswire articles from Hong Kong	CH 55-79
• TREC 2001	English-Arabic: Arabic newswire from *Agence France Presse*	1-25
• TREC2002	English-Arabic: Arabic newswire from *Agence France Presse*	26-75

History

- NTCIR (Japan, NII) (1999 -)
  - Asian languages (CJK) + English
  - Patent retrieval, blogs, Evaluation methodology, ..
- CLEF (Europe) (2000 -)
  - European languages
  - Image retrieval, Wikipedia, …
- SEWM: Chinese IR (2004 -)
- FIRE: IR in Indian languages (2008 -)
- Russian, …

- Search engines
  - Yahoo!: 2006, French/German->German/French, English, Spanish, Italian
  - Google: 2007, Query translation, Translation of retrieved documents
    2012 – fully integrated

CLIR problem

- English query → Chinese document

- Does the Chinese document contain relevant information?

- Readability: In many cases, the translation of retrieved documents into the language of the query is still necessary (goal of machine translation)

Why is it necessary to do CLIR?

- When the relevant information only exists in another language
  - Local information
  - Information about local companies, events, …
  - Patents
- When there is not enough relevant information in the given language
  - Recall-oriented search
- When one can read several languages
  - Avoid submitting multiple queries in different languages
- When the information is language-independent
  - Image
  - Programs, formal specification, …

Strategies for CLIR

- Translate the query
- Translate the documents

- Translate both query and documents into a third language (pivot language)
  - (Related) Transitive translation English-> Chinese
  - Less effective than direct translation
  - Used only when direct translation is impossible

9

Query translation vs. document translation

- Query is short: less contextual information available for translation
- Query translation is more flexible: the user indicates the language(s) of interest, and there is less to be translated

- Document translation: richer context
- More to translate
- Impossible to predetermine in which languages the document should be translated

- Generally, comparable effectiveness

10

How to translate

1. Machine Translation (MT)
2. Bilingual dictionaries, thesauri, lexical resources, …
3. Parallel or comparable corpora

11

Recall: Basics on IR

- What is IR problem
- Basic operations in IR systems
- Models

12

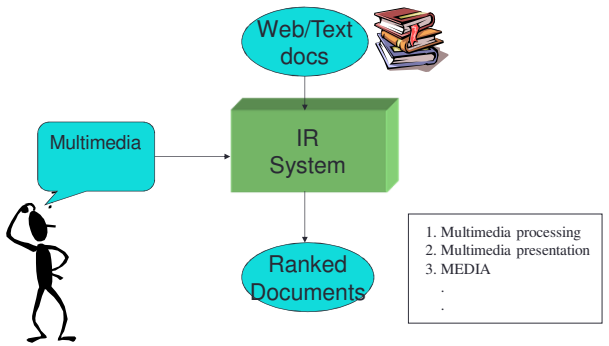
What is IR?

- IR aims at retrieving *relevant* documents from a large document *collection* for a user's *information need*

- Information need → query
- Document collection: a static set of documents
- Relevance: the document contains desired information of the user

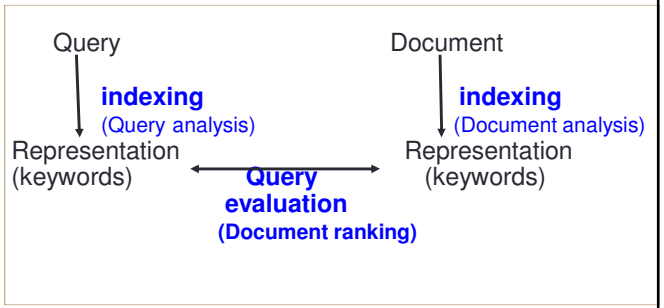
13

System overview



14

Simplified IR architecture



15

Traditional document and query representation

- Using keywords (terms)
- A term is independent from any other term
- Term ←→ Meaning

- In NL, a term (word) may mean the same thing as another term (synonymy)
- It may mean different things in different contexts (polysemy)

16

Problems with Keywords

- Relevant documents may use similar terms
  - "car" vs. "auto"
  - "UK" vs. "England"
  - "computer" vs. "computing"
- Can find irrelevant documents because of the ambiguity of terms.
  - " Mobile " (phone vs. computer vs. qualification)
  - "Apple" (generic vs. the company)
- Can be annoyed by meaningless words
  - "of", "the", "的", …

17

Stopwords / Stoplist

- function words do not bear useful information for IR
  of, in, about, with, I, although, …
- Stoplist: contain stopwords, not to be used as index
  - Prepositions: of, in, …
  - Articles: the, a, …
  - Pronouns: it, me, …
  - Some adverbs and adjectives: great,
  - Some frequent words: document, …
- The removal of stopwords usually improves slightly IR effectiveness
- A few "standard" stoplists are commonly used.

18

Stemming

- Different word forms may bear similar meaning (e.g. compute vs. computing)
- Stemming:
  - Removing some endings of word

computer
compute
computes
computing
computed
computation

→ **comput** (Internal representation)

19

Porter algorithm

(Porter, M.F., 1980, An algorithm for suffix stripping, *Program*, **14**(3) :130-137)

- Step 1: plurals and past participles
  - SSES -> SS — caresses -> caress
  - (*v*) ING -> — motoring -> motor
- Step 2: adj->n, n->v, n->adj, …
  - (m>0) OUSNESS -> OUS — callousness -> callous
  - (m>0) ATIONAL -> ATE — relational -> relate
- Step 3:
  - (m>0) ICATE -> IC — triplicate -> triplic
- Step 4:
  - (m>1) AL -> — revival -> reviv
  - (m>1) ANCE -> — allowance -> allow
- Step 5:
  - (m>1) E -> — probate -> probat
  - (m > 1 and *d and *L) -> single letter — controll -> control

20

Common Preprocessing Steps

- Strip unwanted characters (e.g. punctuation, numbers, etc.).
- Break into tokens (keywords) on whitespace.
- Analyse structure
- Stem tokens into "root" words
  - computational → comput
- Remove common stopwords (e.g. a, the, it, etc.).
- Detect common phrases (possibly using a domain specific dictionary).
- Build inverted index (keyword → list of docs containing it + positions).

21

Retrieval Models

- A retrieval model specifies the details of:
  - Document representation
  - Query representation
  - Retrieval function
- Determines a notion of relevance.

22

Classes of Retrieval Models

- Boolean models (set theoretic)
  - Extended Boolean
- Vector space models (statistical/algebraic)
  - Generalized VSM
  - Latent Semantic Indexing
- Probabilistic models
- Language models

23

Boolean Model

- A document is represented as a set of keywords.
- Queries are Boolean expressions of keywords, connected by AND, OR, and NOT, including the use of brackets to indicate scope.
  - [[Rio & Brazil] | [Hilo & Hawaii]] & hotel & !Hilton]
- Output: Document is relevant or not. No partial matches or ranking.

24

Vector space model

- Assumption: Each term corresponds to a dimension in a vector space
- Vector space = all the keywords encountered in the collection

  $<t_1, t_2, t_3, \ldots, t_n>$
- Document

  $D = < a_1, a_2, a_3, \ldots, a_n>$

  a_i = weight of t_i in D
- Query

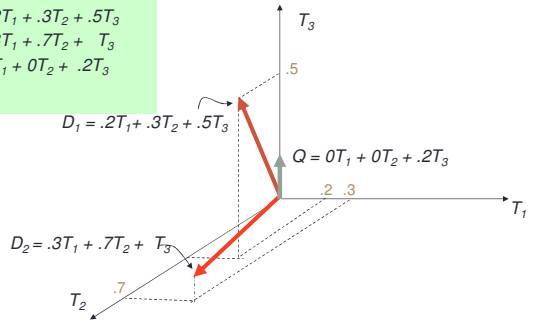
  $Q = < b_1, b_2, b_3, \ldots, b_n>$

  b_i = weight of t_i in Q
- R(D,Q) = Sim(D,Q)

25

Graphic Representation

Example:
$D_1 = .2T_1 + .3T_2 + .5T_3$
$D_2 = .3T_1 + .7T_2 + T_3$
$Q = 0T_1 + 0T_2 + .2T_3$



26

Term Weighting

- Term frequency (TF): the more a term appears in a document, the more it is important
- Terms that appear in many *different* documents are *less* indicative of overall topic (less discriminant)
  - df_i = document frequency of term i
    = number of documents containing term i
  - idf_i = inverse document frequency of term i,
    = $\log_2 (N/df_i)$
    (N: total number of documents)
- A typical combined term importance indicator is *tf-idf weighting*

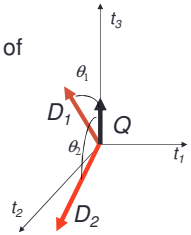
$$w_{ij} = tf_{ij} idf_i = tf_{ij} \log_2 (N/df_i)$$

27

Similarity Measures

- Inner product
- Cosine similarity measures the cosine of the angle between two vectors.

$$\cos(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} (w_{ij} \times w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \times \sum_{i=1}^{t} w_{iq}^2}}$$



Probabilistic model

- Given D, estimate P(R|D,Q) and P(NR|D,Q)

$$P(R|Q,D) = \frac{P(D|R,Q) * P(R|Q)}{P(D|Q)}$$

- P(D|Q), P(R|Q) constant

→ P(R|Q,D)∝ P(D|R)

- D = {t_1=x_1, t_2=x_2, …}

$$x_i = \begin{cases} 1 \text{ if present} \\ 0 \text{ if absent} \end{cases}$$

$$P(D|R,Q) = \prod_{(t_i=x_i) \in D} P(t_i = x_i | R_Q)$$

$$P(D|NR,Q) = \prod_{(t_i=x_i) \in D} P(t_i = x_i | NR_Q)$$

$$Odd(D) = \log \frac{P(D|R,Q)}{P(D|NR,Q)}$$

How to estimate $P(t_i = x_i \mid R_Q)$

- A set of *N* relevant and irrelevant samples:

$$P(t_i = 1 \mid R_Q) = \frac{r_i}{R_i}$$

$$P(t_i = 1 \mid NR_Q) = \frac{n_i - r_i}{N - R_i}$$

r_i	$n_i - r_i$	n_i Doc. with t_i
$R_i - r_i$	$N - R_i - n + r_i$	$N - n_i$ Doc. without t_i
R_i Rel. doc	$N - R_i$ Irrel.doc.	N Samples

BM25

$$Score(D,Q) = \sum_{t \in Q} w \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} + k_2 \mid Q \mid \frac{avdl - dl}{avdl + dl}$$

$$K = k_1((1-b) + b\frac{dl}{avdl - dl})$$

TF factors

Doc. length normalization

- k_1, k_2, k_3, b: parameters determined empirically
- tf: term frequency
- qtf: query term frequency
- dl: document length
- avdl: average document length

Some additional factors

- Smoothing (Robertson-Sparck-Jones formula)

$$Odd(D) = \sum_{t_i} x_i \log \frac{(r_i + 0.5)(N - R_i - n_i + r_i + 0.5)}{(R_i - r_i + 0.5)(n_i - r_i + 0.5)} = \sum_{t_i \in D} x_i w_i$$

- When no sample is available:
  P(t=1|R)=0.5,
  $P(t=1|NR)=(n_i+0.5)/(N+0.5) \approx n_i/N$

- Use relevance feedback to get more samples for more precise estimation (usually unavailable)

Language models

Question: Is the document likelihood increased when a query is submitted?

$$LR(D,Q) = \frac{P(D \mid Q)}{P(D)} = \frac{P(Q \mid D)}{P(Q)}$$

(Is the query likelihood increased when D is retrieved?)

- P(Q|D) calculated with $P(Q|M_D)$
- P(Q) estimated as $P(Q|M_C)$

$$Score(Q,D) = \log \frac{P(Q \mid M_D)}{P(Q \mid M_C)}$$

Divergence between M_D and M_Q

$$Score(Q,D) = \sum_{i=1}^{n} tf(q_i,Q) * \log \frac{P(q_i \mid M_D)}{P(q_i \mid M_C)}$$

$$\propto \sum_{i=1}^{n} P(q_i \mid M_Q) * \log \frac{P(q_i \mid M_D)}{P(q_i \mid M_C)}$$

$$\propto \sum_{i=1}^{n} P(q_i \mid M_Q) * \log P(q_i \mid M_D)$$

Query model: Max. likelihood

$$P(q_i \mid M_Q) = \frac{tf(q_i,Q)}{|Q|}$$

Document model: Needs smoothing

Smoothing

- In IR, combine doc. with corpus
  - Interpolation

$$P(w_i \mid M_D) = \lambda P_{ML}(w_i \mid M_D) + (1-\lambda) P_{ML}(w_i \mid M_C)$$

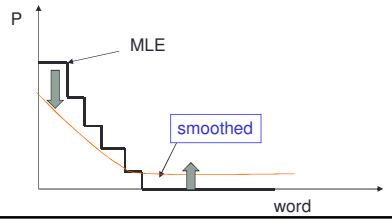
  - Dirichlet

$$P_{Dir}(w_i \mid M_D) = \frac{tf(w_i,D) + \mu P_{ML}(w_i \mid M_C)}{|D| + \mu}$$

34

Document model Smoothing

- Zero probability (log 0)
- Goal: assign a low probability to words or n-grams not observed in the training corpus



36

Query expansion

- Original query *Q*
- Determine some expansion terms (e.g. synonyms, related terms)
- Add the expansion terms into *Q* → *Q'*
- Use *Q'* to retrieve documents

- *Q'* is usually better than *Q*

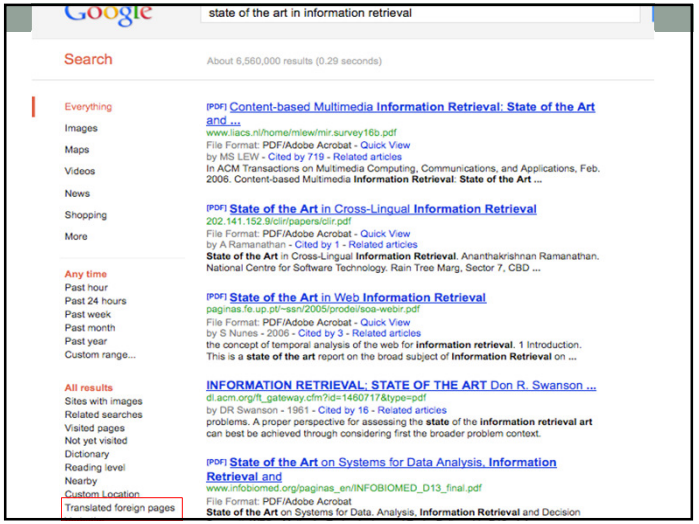
- In language model:

$$P(w_i \mid M_Q) = \lambda P_{ML}(w_i \mid M_Q) + (1-\lambda) \sum_{w_j} P(w_i \mid w_j) P_{ML}(w_j \mid M_Q)$$

37

Back to CLIR: How to translate

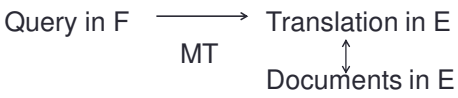
1. Machine Translation (MT)
2. Bilingual dictionaries, thesauri, lexical resources, …
3. Parallel texts: translated texts



38

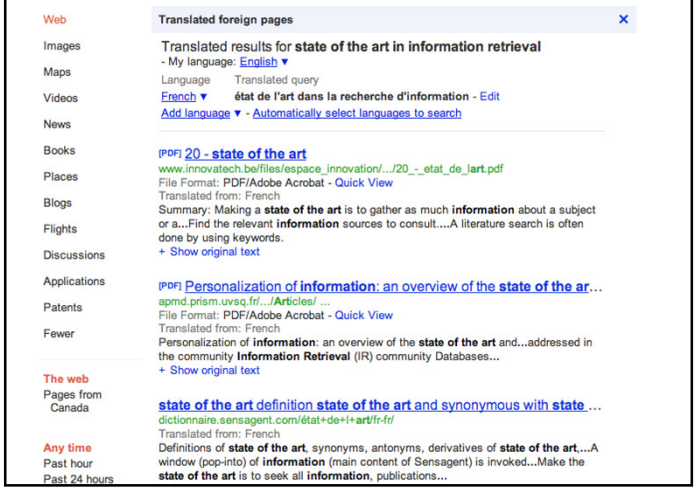
Approach 1: Using MT

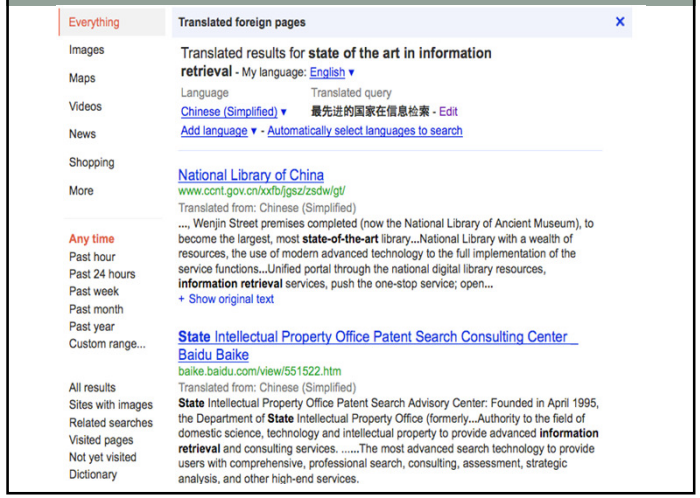
- Seems to be the ideal tool for CLIR and MLIR (if the translation quality is high)

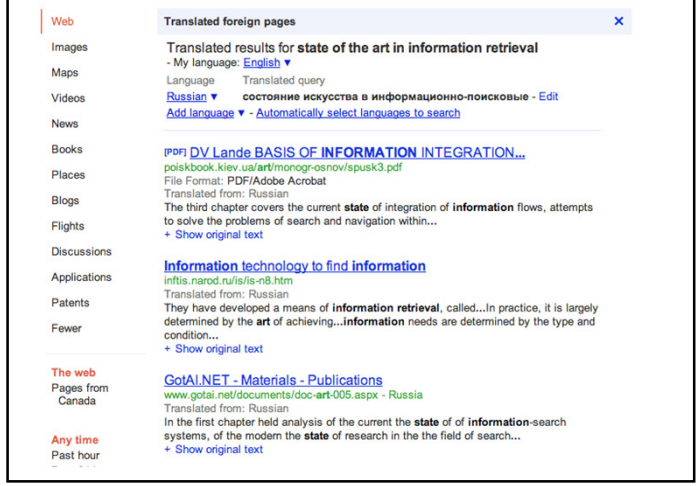


- Typical effectiveness: 80-100% of the monolingual effectiveness
- Problems:
  - Quality
  - Availability

40







Query translation using MT

- Query translation: to make query comparable to documents
- Similarities with MT
  - Similar translation problems
  - Similar methods can be used
- A good MT → good CLIR effectiveness

Differences between query translation and MT

- Short queries (2-3 words): HD video recording
- Flexible syntax: video HD recording, recording HD…
- Goal: help find relevant documents, not to make the translated query readable by users
  - Less strict translation: The "translation" can be by related words (same/related topics)
    → Query expansion effect
  - Not limited to one translation (e.g. potato → pomme de terre /patate, 土豆 / 马铃薯)
- Not only translation
  - weight = correctness of translation + Utility for IR
  - E.g. Organic food
    Translation → 有机食品 (organic food)
    Utility for RI → 绿色食品 (green food)

45

Translation problems

西点 : western dessert / West Point

46

Translation problems

Caution: slippery steps

47

Problems of MT for IR

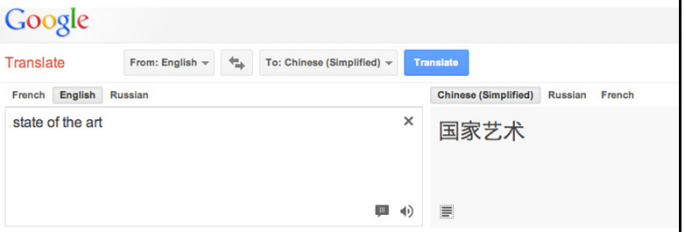
Exit

48

Translation problems

49

Translation problems



Translated as “art of the state”

50

Exemples

	Systran	Google
• 1. drug traffic		
	trafic de stupéfiants	trafic de stupéfiants
	毒品交易	毒品贩运
• 2. drug insurance:		
	assurance de drogue	d'assurance médicaments
	药物保险	药物保险
• 3. drug research:		
	recehrche de drogue	la recherche sur les drogues
	药物研究	药物研究

51

Examples

	Systran	Google
4. drug for treatment of Friedreich’s ataxia:		
	drogue pour le traitement de l'ataxie de Friedreich	médicament pour le traitement de l'Ataxie de Friedreich
	Friedreich的不整齐的治疗的药物	药物治疗弗里德的共济失调
5. drug control:		
	commande de drogue	contrôle des drogues
	药物管制	药物管制
• 6. drug production:		
	production de drogue	la production de drogues
	药物生产	药物生产

52

Approach 2: Using bilingual dictionaries

- Unavailability of high-quality MT systems for many language pairs
- MT systems can often be used as a black box that is difficult to adapt to IR task
  - MT produces one best translation
  - IR needs translation alternatives
- Bilingual dictionary:
  - A non-expensive alternative
  - Usually available

53

Approach 2: Using bilingual dictionaries

- General form of dict. (e.g. Freedict)

access:	attaque, accéder, intelligence, entrée, accès
academic:	étudiant, académique
branch:	filiale, succursale, spécialité, branche
data:	données, matériau, data

- LDC English-Chinese
  - AIDS /艾滋病/爱滋病/
  - data /材料/资料/事实/数据/基准/
  - prevention /阻碍/防止/妨碍/预防/预防法/
  - problem /问题/难题/疑问/习题/作图题/将军/课题/困难/难/题是/
  - structure /构造/构成/结构/组织/化学构造/石理/纹路/构造物/建筑物/建造/物/

54

Basic methods

- Use all the translation terms
  - data /材料/资料/事实/数据/基准/
  - structure /构造/构成/结构/组织/化学构造/石理/纹路/构造物/建筑物/建造/物/
  - Introduce noise
  - Implicitly, the term with more translations is assigned higher importance
- Use the first (or most frequent) translation
  - Does not cover other translation alternatives
  - Not always an appropriate choice
- Generally: 50-60% of monolingual IR

- Problems of dictionary
  - Coverage (unknown words, unknown translations)
  - [Xu and Weischedel 2005] tested the impact of dictionary coverage on CLIR (En-Ch)
  - The effectiveness increases till 10 000 entries

55

Translate the query as a whole

- Phrase translation [Ballesteros and Croft, 1996, 1997]
  - base de données: database
  - pomme de terre: potato
  - First translate phrases
  - Then the remaining words

- Best global translation for the whole query

1. Candidates:
   For each query word
   - Determine all the possible translations (through a dictionary)
2. Selection
   select the set of translation words that produce the highest ***cohesion***

56

Cohesion

- Cohesion ~ frequency of two translation words together
  E.g.
  - data: données, matériau, data
  - access: attaque, accéder, intelligence, entrée, accès

  (accès, données) 152 *
  (accéder, données) 31
  (données, entrée) 21
  (entrée, matériau) 3
  …

  - Freq. from a document collection or from the Web (Grefenstette 99)
  - (Gao, Nie et al. 2001) (Liu and Jin 2005)(Seo et al. 2005)…

$$\arg\max_{T_Q} Cohesion(T_Q) = \arg\max_{T_Q} \sum_{t_i \in Q} \sum_{t_j \in Q \wedge t_j \neq t_i} sim(T_{t_i}, T_{t_j})$$

  - sim: co-occurrence, mutual information, statistical dependence

- Improved effectiveness (80-100% of monolingual IR)

57

Approach 3: using parallel texts

- Training a translation model (IBM 1) $t(t_j|s_i)$
- Principle:
  - Principle: The more s_j appears in parallel texts of t_i, the higher $t(t_j|s_i)$.

$\dots s_i \dots \longleftrightarrow \dots t_j \dots$

$\dots s_i \dots \longleftrightarrow \dots t_j \dots$

$\dots s_i \dots \longleftrightarrow \dots t_k \dots$

58

Utilization of TM in CLIR

- Query: a set of source words
- Each source word: a set of weighted target words
  - some filtering on translations: stopwords, prob. threshold, number of translations, …
- All the target words ≈ query "translation"
  - Common approach: a set of weighted translation terms in a bag of words
- Using "translated" bag of words in monolingual IR

59

Simple utilization

- Directly use the probability of a word translation

$$P(f|Q_E) = \sum_{e \in Q_E} t(f|e) \cdot P(e|Q_E)$$

- One should also take into account the discriminant power of the translation (IDF) (better in VSM)

$$w(f, Q_E) = \sum_{e \in Q_E} t(f|e) \cdot P(e|Q_E) \cdot \log \frac{|C_F|}{n_f}$$

60

example

Query #3 What measures are being taken to stem international drug traffic?

médicament=0.110892
mesure=0.091091
international=0.086505
trafic=0.052353
drogue=0.041383
découler=0.024199
circulation=0.019576
pharmaceutique=0.018728
pouvoir=0.013451
prendre=0.012588
extérieur=0.011669
passer=0.007799
demander=0.007422
endiguer=0.006685
nouveau=0.006016
stupéfiant=0.005265
produit=0.004789

- Multiple translations,
- but ambiguity is kept
- Difficult to distinguish usual translation words from coincidental translations (e.g. pouvoir)
- Unknown word in target language

IBM1 + dictionary: a simple combination in VSM

- Increase the weight of usual translation words in the dictionary (TREC-6)

	Number of translation words					
Default probability	10	20	30	40	50	100
0.005	0.2671	0.2787	0.2812	0.2813	0.2829	0.2671
0.01	0.2755	0.2873	0.2891	0.2896	0.2906	0.2742
0.02	0.2873	0.2959	0.2962	0.2967	**0.2985**	0.2825
0.03	0.2811	0.2906	0.2898	0.2897	0.2904	0.2744
0.04	0.2751	0.2842	0.2827	0.2826	0.2831	0.2683
0.05	0.2687	0.2761	0.2729	0.2729	0.2730	0.2578

Number *N* of translation words	MAP (% monolingual IR)
10	0.2546 (68.24%)
20	0.2635 (70.62%)
30	0.2660 (71.30%)
40	0.2664 (71.40%)
50	0.2671 (71.59%)
100	0.2506 (67.14%)

- MAP-mono = 0.3731
- MAP-LOGOS = 0.2866 (76.8%), MAP-Systran = 0.2763 (74.1%)

Principle of model training

- $t(t_j|s_i)$ is estimated from a parallel training corpus, aligned into parallel sentences
- IBM models 1, 2, 3, …
- Process:
  - Input = two sets of parallel texts
  - Sentence alignment A: $S_k \Leftrightarrow T_l$ (bitext)
  - Initial probability assignment: $t(t_j|s_i,A)$
  - Expectation Maximization (EM): $t(t_j|s_i,A)$
  - Final result: $t(t_j|s_i) = t(t_j|s_i,A)$

Integrating translation in language model (Kraaij et al. 2003)

- The problem of CLIR:

$$Score(D,Q) = \sum_{t_i \in V} P(t_i \mid M_Q) \log P(t_i \mid M_D)$$

- Query translation (QT)

$$P(t_i \mid M_{Q_s}) = \sum_{s_j \in V_s} P(t_i \mid s_j, M_{Q_s}^{ML}) P(s_j \mid M_{Q_s}^{ML})$$

$$\approx \sum_{s_j \in V_s} t(t_i \mid s_j) P(s_j \mid M_{Q_s}^{ML})$$

- Document translation (DT)

$$P(s_i \mid M_{D_t}) = \sum_{t_j \in V_t} P(s_i \mid t_j, M_{D_t}) P(t_j \mid M_{D_t})$$

$$\approx \sum_{t_j \in V_t} t(s_i \mid t_j) P(t_j \mid M_{D_t})$$

Results (CLEF 2000-2002)

Run	EN-FR	FR-EN	EN-IT	IT-EN
Mono	0.4233	0.4705	0.4542	0.4705
MT	0.3478	0.4043	0.3060	0.3249
QT	0.3878	**0.4194**	0.3519	**0.3678**
DT	**0.3909**	0.4073	**0.3728**	0.3547

- Translation model (IBM 1) trained on a web collection
- TM can outperform MT (Systran)

65

Details on translation model training on a parallel corpus

- Sentence alignment
  - Align a sentence in the source language to its translation(s) in the target language
- Translation model
  - Extract translation relationships
  - Various models (assumptions)

66

Sentence alignment

- Assumption:
  - The order of sentences in two parallel texts is similar
  - A sentence and its translation have similar length (length-based alignment, e.g. Gale & Church)

$$D(i,j) = \min \begin{cases} D(i,j-1) + d_1 \\ D(i-1,j) + d_2 \\ D(i-1,j-1) + d_3 \\ D(i-1,j-2) + d_4 \\ D(i-2,j-1) + d_5 \\ D(i-2,j-2) + d_6 \end{cases}$$

d_i: distance for different patterns (0-1, 1-1, …)

  - A translation contains some “known” translation words, or cognates (e.g. Simard et al. 93)

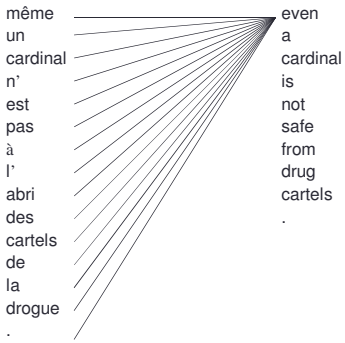
67

Example of aligned sentences (Canadian Hansards)

Débat L'intelligence artificielle	Artificial intelligence A ebate	2-2
Depuis 35 ans, les spécialistes d'intelligence artificielle cherchent à construire des machines pensantes. Leurs avancées et leurs insuccès alternent curieusement.	Attempts to produce thinking machines have met during the past 35 years with a curious mix of progress and failure.	2-1
	Two further points are important.	0-1
Les symboles et les programmes sont des notions purement abstraites.	First, symbols and programs are purely abstract notions.	1-1

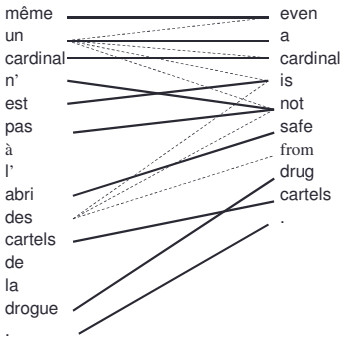
68

TM training: Initial probability assignment $t(t_j|s_i, A)$



69

TM training:
Application of EM: $t(t_j|s_i, A)$

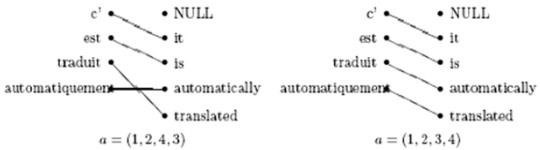


Word alignment for one sentence pair

Source sentence in training: **e** = $e_1, \ldots e_l$ (+NULL)

Target sentence in training: **f** = $f_1, \ldots f_m$

Only consider alignments in which each target word (or position j) is aligned to a source word (of position a_j)



The set of all the possible word alignments: A(**e,f**)

70

IBM models (Brown et al.)

- IBM 1: does not consider positional information and sentence length
- IBM 2: considers sentence length and word position
- IBM 3, 4, 5: fertility in translation

- For CLIR, IBM 1 seems to correspond to the current (bag-of-words) approaches to IR.

72

General formula

$$P(F = \mathbf{f} \mid E = \mathbf{e}) = P(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a} \in A(\mathbf{e},\mathbf{f})} P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

$$\mathbf{a} = (a_1, \ldots, a_m) \text{ with } a_i \in [0, l] \; \forall i \in [1, m], l = |\mathbf{e}|, m = |\mathbf{f}|$$

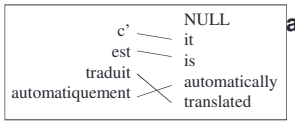
$$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = P(m \mid \mathbf{e}) \prod_{j=1}^{m} P(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) P(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e})$$

Prob. that **e** is translated into a sentence of length m

Prob. that j-th target is aligned with a_j-th source word

Prob. to produce the word f_j at position j

Example



a = (1,2,4,3)

$P(m \mid \mathbf{e}) \prod_{j=1}^{m} P(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) P(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e})$

p(c'est traduit automatiquement, **a**|it is automatically translated) =

$P(m=4 \mid \mathbf{e}=$ it is automatically translated$)$ □

$P_a(a_1=1 \mid m=4, \mathbf{e})$ □ $P_t(f_1=$ c' $\mid a_1=1, m=4, \mathbf{e})$ □

$P_a(a_2=2 \mid a_1=1, f_1=$ c', $m=4, \mathbf{e})$ □

$P_t(f_2=$ est $\mid a_1^2=(1,2), f_1=$ c', $m=4, \mathbf{e})$ □

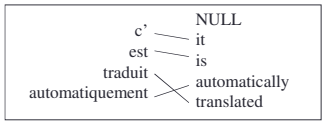
$P_a(a_3=4 \mid a_1^2=(1,2), f_1^2=$ c' est, $m=4, \mathbf{e})$ □

$P_t(f_3=$ traduit $\mid a_1^3=(1,2,4), f_1^2=$ c'est, $m=4, \mathbf{e})$ □

$P_a(a_4=3 \mid a_1^3=(1,2,4), f_1^3=$ c' est traduit, $m=4, \mathbf{e})$ □

$P_t(f_4=$ automatiquement $\mid a_1^4=(1,2,4,3), f_1^3=$ c'est traduit, $m=4, \mathbf{e})$

Example Model 1



a = (1,2,4,3)

Assume :

t(c'| it) = 0.2,

t(est | is) = 0.7,

t(traduit | translated) = 0.45

t(automatiquement | automatically) = 0.8

p(c'est traduit automatiquement, a | it is automatically translated)

$$= \frac{\varepsilon}{5^4} \times 0.2 \times 0.7 \times 0.45 \times 0.8 = \varepsilon \times 8.064 \times 10^{-5}$$

IBM model 1

- Simplifications

$P(m \mid \mathbf{e}) = \varepsilon$ — Any length generation is equally probable – a constant

$P(a_j \mid a_1^{j-1} f_1^{j-1}, m, \mathbf{e}) = p(a_j \mid l) = 1/(l+1)$ — Position alignment is uniformly distributed

$P(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e}) = p(f_j \mid e_{a_j}) = t(f_j \mid e_{a_j})$ — Context-independent word translation

- the model becomes (for one sentence alignment **a**)

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \varepsilon \times \prod_{j=1}^{m} \frac{t(f_j \mid e_{a_j})}{l+1}$$

$$= \frac{\varepsilon}{(l+1)^m} \times \prod_{j=1}^{m} t(f_j \mid e_{a_j})$$

Sum up all the alignments

$$p(\mathbf{f} \mid \mathbf{e}) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j \mid e_{a_j})$$

$$= \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j \mid e_i)$$

- Problem:
  We want to optimize $t(f_j \mid e_i)$ so as to maximize the likelihood of the given sentence alignments
- Solution: Using EM

Parameter estimation

1. An initial value for $t(f|e)$ (f, e are words)
2. Compute the count $c(f|e, \mathbf{e}^{(s)}, \mathbf{f}^{(s)})$ of word alignment e-f in the pair of sentences ($\mathbf{e}^{(s)}, \mathbf{f}^{(s)}$) (E-step)

$$c(f|e;\mathbf{e}^{(s)},\mathbf{f}^{(s)}) = \sum_{\mathbf{a}} p(\mathbf{a}|\mathbf{e}^{(s)},\mathbf{f}^{(s)}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j})$$

$$= \frac{t(f|e)}{\sum_{i=0}^{l} t(f|e_i)} \underbrace{\sum_{j=1}^{m} \delta(f, f_j)}_{\text{Count of } f \text{ in } \mathbf{f}} \underbrace{\sum_{i=0}^{l} \delta(e, e_i)}_{\text{Count of } e \text{ in } \mathbf{e}}$$

3. Maximization (M-step)

$$\lambda_e = \sum_f \sum_{s=1}^{S} c(f|e;\mathbf{e}^{(s)},\mathbf{f}^{(s)}) \quad \text{(normalization factor)}$$

$$t(f|e) = \lambda_e^{-1} \sum_{s=1}^{S} c(f|e;\mathbf{e}^{(s)},\mathbf{f}^{(s)})$$

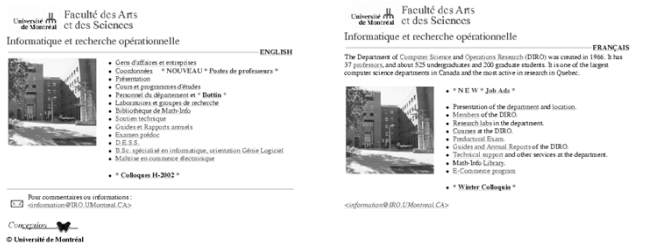
4. Loop on 2-3

Problem of parallel texts

- Only a few large parallel corpora
  - e.g. Canadian Hansards, EU parliament, Hong Kong Hansards, UN documents, …
- Many languages are not covered
- Is it possible to extract parallel texts from the Web?
  - STRANDS
  - PTMiner

An example of "parallel" pages

http://www.iro.umontreal.ca/index.html http://www.iro.umontreal.ca/index-english.html



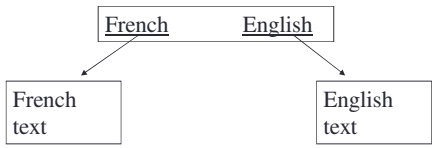
STRANDS [Resnik 98]

- Assumption:

  If - A Web page contains 2 pointers

  - The anchor text of each pointer identifies a language

  Then The two pages referenced are "parallel"



81

PTMiner (Nie & Chen 99)

- Candidate Site Selection

  By sending queries to AltaVista, find the Web sites that may contain parallel text.

- File Name Fetching

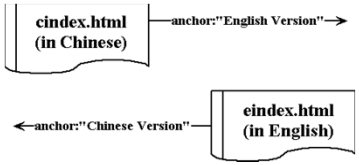
  For each site, fetching all the file names that are indexed by search engines. Use host crawler to thoroughly retrieve file names from each site.

- Pair Scanning

  From the file names fetched, scan for pairs that satisfy the common naming rules.

82

Candidate Sites Searching



- Assumption: A candidate site contains at least one such Web page referencing another language.
- Take advantage of existing search engines (AltaVista)

83

File Name Fetching

- Initial set of files (seeds) from a candidate site:

  host:www.info.gov.hk

- Breadth-first exploration from the seeds to discover other documents from the sites

84

Pair Scanning

- Naming examples:

  index.html v.s. index_f.html

  /english/index.html v.s. /french/index.html

- General idea:

  parallel Web pages = Similar URLs at the difference of a tag identifying a language

85

Mining Results (several years ago)

- French-English
  - Exploration of 30% of 5,474 candidate sites
  - 14,198 pairs of parallel pages
  - 135 MB French texts and 118 MB English texts
- Chinese-English
  - 196 candidate sites
  - 14,820 pairs of parallel pages
  - 117.2M Chinese texts and 136.5M English texts
- Several other languages I-E, G-E, D-E, …

86

CLIR results: F-E

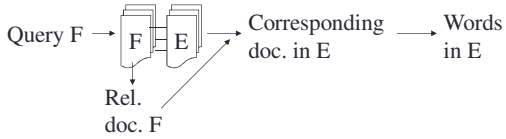
	F-E (Trec6)	F-E (Trec7)	E-F (Trec6)	E-F (Trec7)
Monolingual	0.2865	0.3202	0.3686	0.2764
Hansard TM	0.2166 (74.8%)	0.3124 (97.6%)	0.2501 (67.9%)	0.2587 (93.6%)
Web TM	0.2389 (82.5%)	0.3146 (98.3%)	0.2504 (67.9%)	0.2289 (82.8%)

- Web TM comparable to Hansard TM
- Parallel texts for CLIR can tolerate more noise

87

Alternative methods

- using parallel texts for pseudo-relevance feedback [Yang et al. 99]
  - Given a query in F
  - Find relevant documents in the parallel corpus
  - Extract keywords from their parallel documents, and consider them as a query translation



88

Alternative methods – LSI [Dumais et al. 97]

- Monolingual LSI :
  - Create a latent semantic space (using SVD)
    - Each dimension represents a combination of initial dimensions (terms, documents)
  - Comparison of document-query in the new space
- Bilingual LSI :
  - Create a latent semantic space for both languages on a parallel corpus
    - Concatenate two parallel texts together
    - Convert terms in both languages into the semantic space
- Problems:
  - The dimensions in the latent space are determined to minimize some representational error – may be different from translational error
  - Coverage of terms by the parallel corpus
  - Complexity in creating the semantic space
- Effectiveness – usually lower than using a translation model

89

Explicit Semantic Analysis (ESA)
[Gabrilovich et al. 07, Spitkovsky et al. 12]

- Assume that each Wikipedia article corresponds to one explicit representation dimension
- A term → association with an article (tf*idf or conditional probability) based on text body or anchor text
- Document ranking: compare the document and the query representations in the ESA space

- CLIR:
  - Using cross-lingual links between Wikipedia articles

- Possible problems:
  - Completeness of ESA space
  - Representation granularity

90

Summary on using document-level term relations

- Pseudo-RF, LSI, ESA
- Assumption: Terms occurring in the same text (or parallel text) are related
- (Translation) relations between terms are coarse-grained
  - Related to the same topics
  - Query "translation" by topic-related terms
- Usually lower retrieval effectiveness than explicit translation relations
  - May be suitable for comparable texts
  - Do not exploit fully parallel texts

91

Using a comparable corpus

- Comparable: articles about the same topic
  - E.g. News articles on the same day about an event

- Impossible to train a translation model
- Estimate cross-lingual similarity (less precise than translation)
  - Similar methods to co-occurrence analysis
  - Conditional probability, mutual information, …
- Less effective than using a parallel corpus
- To be used only when there is no parallel corpus, or to complement a dictionary or parallel corpus
  - Helpful to further expand the translation by a dictionary, MT or TM

92

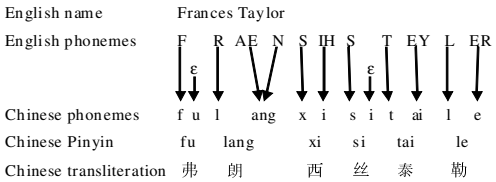
Other problems – unknown words

- Proper names ('Vladmir Ivanov' in Chinese?)
- New technical terms ('Latent Semantic Analysis' in Chinese?)

- Possible solutions
  - Transliteration
  - Mining the web

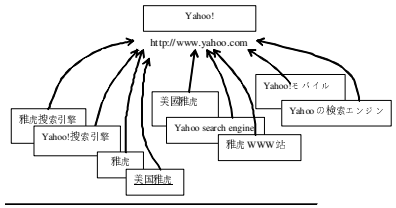
Transliteration (between languages of different alphabets)

- Translate a name phonetically
  - Generate the pronunciation of the name
  - Transform the sounds into the target language sounds
  - Generate the characters to represent the sounds



Mining the web (1)

- A site is referred to by several pages with different anchor texts in different languages
- Anchor texts as parallel texts
- Useful for the translation of organizations (故宫博物馆 – National Museum)



Mining the web (2)

- Some "monolingual" texts may contain translations
  - 现在网上最热的词条，就是这个“Barack Obama”，巴拉克·欧巴马（巴拉克·奥巴马）。
  - 这就诞生了潜语义索引(Latent Semantic Indexing) …
  - Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais *Support Vector Machine*, SVM) sont …
- Procedure
  - Using templates to identify potential candidates:
    Source-name (Target-name)
    Source-name, Target-name
    …
  - Segmentation (what segment may be appropriate?)
  - Statistical analysis
- May be used to complete an existing dictionary

Mining the Web (3)

- Wikipedia – a rich source for translations
- Links between articles in different languages
  - Translation of concepts (Wiki titles) and named entities (e.g. proper names) through cross-language links
  - Linked articles as comparable texts → term similarity

97

Some other improvement means

- Pre- and post-translation expansion
  - Query expansion before the translation using a source-language collection
  - Query expansion after the translation using a target-language collection (traditional PRF)
- Fuzzy matching between similar languages
  - information - información – informazione
  - ~cognate
  - Matching n-grams (e.g. 4-grams)
  - Transformation using rules (*konvektio -> convection*)
- Combining translations using different tools
  - Several MT systems, several dictionaries
  - MT + TM + dictionary
  - Parallel texts + comparable texts
  - …

98

Structured query

- Traditional method: all the translation terms in a bag of words
  - data /材料/资料/事实/数据/基准/
  - structure /构造/构成/结构/组织/化学构造/石理/纹路/构造物/建筑物/建造/物/
- Consider all translations for the same source word as synonyms
  - #syn(材料 资料 事实 数据 基准) #syn(构造 构成 …)
  - sum up the occurrences of all the synonyms in a document (v.s. sum up the log probabilities without #syn)
- Pirkola: structured query > bag-of-words query
- Probabilistic structured query (#wsyn): add weights to synonyms

99

Current state

- Effectiveness of CLIR
  - Between European languages ~90-100% monolingual
  - Between European and Asian languages ~ 80-100%
- A usable quality
- One usually needs translation of the retrieved documents
- The use of CLIR by Web users is still limited / Tools for CLIR are limited
  - To be increased (integration in the main search engines)

100

Remaining problems

- Current approaches :
  - CLIR= translation + monolingual IR
    - E.g. Using MT as a black box + monolingual IR
  - The resources and tools are usually developed for MT, not for CLIR
    - MT: create a readable sentence
    - CLIR: retrieving relevant documents
- Problems of translation selection
  - MT: Select one best translation → multiple translations
  - Phrases in MT: consecutive words
    - But dependent words do not always form a phrase
    - "Mixing **drug** cocktails for mental *illness* is still more art than science"
    - →Take into account more flexible dependencies (even proximity)
    - How to train a translation model in such a context?
  - These are not only a problem in CLIR but also in general IR.

The future?

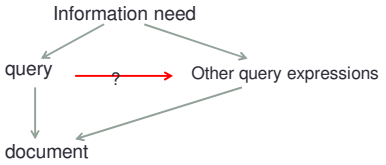
- CLIR≠ translation+ monolingual IR
- Translation as a step in CLIR
  - Translation for IR (not for human readers)
  - Select effective *search* terms (not only good *translation* terms)
  - Some similarities with query expansion
    - Can use similar approaches to query expansion
  - Combine multiple translation possibilities
    - Result diversification
- Document translation: Query-dependent?
- (More in the talk on SMT)

Beyond

- Current CLIR approaches can succeed in crossing the language barrier
  - MT, Dictionary, STM
- Can one use CLIR methods for monolingual (general) IR?
  - Basic idea
  - IBM1 model
  - Phrase translation model and some adaptations

What have we learnt from CLIR?

- The original query is not always the best expression of information need.
- We can generate better or alternative query expressions
  - Query expansion, reformulation, rewriting
- General IR



Basic idea for general IR

- Assumption: Lexical gap between queries and documents
  - Written by searchers and by authors, using different vocabularies
- Translate between query language and document language
- How?
  - Dictionary, thesaurus (~dictionary-based translation)
  - Statistical “translation” models (*)

Translation in monolingual IR

- Translation model (Berger&Lafferty, 99)

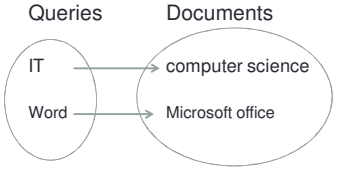
$$P(q_i \mid D) = \sum_{d_j} t(q_i \mid d_j) P(d_j \mid D)$$

$$P(d_i \mid Q) = \sum_{q_j} t(d_i \mid q_j) P(q_j \mid Q)$$

- Used for document/query expansion
- How to obtain word translations?
  - Pseudo-parallel texts: A sentence is "parallel" to itself, to the paragraph, …
  - co-occurrence analysis: relate a term and its context terms
  - A document title is "parallel" to its contents (title language model – Jin 2002)
  - Documents on the same topic are comparable
  - A query is "parallel" to the clicked document (query logs)
  - …

Using query logs

- Query language ≠ document language



- Wen et al, 2001, Cui et al. 2002
  - Term co-occurrences across queries and clicked documents
  - Term similarity
  - Query clustering and query expansion
- Gao et al. 2010
  - Using translation model on query-document title pairs

Query logs

msn web	0.6675749
Webmensseger	0.6621253
msn online	0.6403270
windows web messanger	0.6321526
talking to friends on msn	0.6130790
school msn	0.5994550
msn anywhere	0.5667575
web message msn com	0.5476839
msn messager	0.5313351
hotmail web chat	0.5231608
messenger web version	0.5013624
instant messager msn	0.4[illegible]550409
browser based messenger	0.3814714
im messenger sign in	0.2997275
…	…

Title: msn web messenger

Figure 1: A fragment of the query click field for the page *http://webmessenger.msn.com* [16].

Perplexity results

Order	Body	Anchor	Title	Query
Unigram	13242	4164	3633	1754
Bigram	5567	966	1420	289
Trigram	5381	740	1299	180
4-gram	5785	731	1382	168

Table 2: Perplexity results on test queries, using *n*-gram models with different orders, derived from different data sources.

- Test set
  - 733,147 queries from the May 2009 query log
- Summary
  - Query LM is most predictive of test queries
  - Title is better than Anchor in lower order but is worse in higher order
  - Body is in a different league

Word-based translation model

- Basic model: $P(Q|D) = \prod_{q \in Q} \sum_{w \in D} P(q|w)P(w|D)$
- Mixture model:

$$P_s(q|D) = \alpha P(q|C) + (1-\alpha)P_{mx}(q|D), \text{ where}$$

$$P_{mx}(q|D) = \beta P(q|D) + (1-\beta) \sum_{w \in D} P(q|w)P(w|D)$$

- Learning translation probabilities from clickthrough data
  - IBM Model 1 with EM

$$P(Q|D,\theta) = \frac{\varepsilon}{(I+1)^J} \prod_{q \in Q} \sum_{w \in D} P(q|w,\theta)$$

Results

#	Models	NDCG@1	NDCG@3	NDCG@10	
1	BM25	0.3181	0.3413	0.4045	
2	WTM_M1 (β=1)	0.3202	0.3445	0.4076	No TM
3	WTM_M1	0.3310	0.3566	0.4232	
4	WTM_M1 (β=0)	0.3210	0.3512	0.4211	Only TM

Using phrase translations

- Phrase = sequence of adjacent words
- Words are often ambiguous
- Phrase translations are more precise: natural contextual constraint between words in a phrase
- State of the art in MT: Using phrase translation
- Question:

  Is phrase-based query translation effective in IR?

Phrase translation in IR (Riezler et al. 2008)

Phrase = consecutive words

Determining phrases:

- Word translation model
- Word alignments between parallel sentences
- Phrase = intersection of alignments in both directions

	QE Models	NDCG@1	NDCG@3	NDCG@10
1	NoQE	0.2803	0.3430	0.4199
2	PhraseMT	0.3002	0.3617	0.4363
3	WordModel	0.3117	0.3717	0.4434

- Note: on a different collection

113

Why isn't phrase-model better than word-model?

- Sentence (Riezler et al 2008)

  herbs for chronic constipation

- Words and phrases
  - herbs
  - herbs for
  - herbs for chronic
  - for
  - for chronic
  - for chronic constipation
  - chronic constipation
  - constipation

114

More flexible "phrases" – n-grams for query

Word generation process:

- Select a segmentation S of Q according to $P(S|Q)$,
- Select a query phrase $\mathbf{q}$ according to $P(\mathbf{q}|S)$, and
- Select a translation (i.e., expansion term) w according to $P(w|\mathbf{q})$.

$$P(w|Q) = \sum_{S} \sum_{\mathbf{q} \in S} P(S|Q) P(\mathbf{q}|S) P(w|\mathbf{q})$$

Query	**Title**
instant messenger msn	msn web messenger
instant messenger msn instant-messenger messenger-msn instant-messenger-msn	msn web messenger

115

Comparison

	QE Models	NDCG@1	NDCG@3	NDCG@10
1	NoQE	0.2803	0.3430	0.4199
2	WM	0.3117	0.3717	0.4434
3	PM_2	0.3261	0.3832	0.4522
4	PM_8	0.3263	0.3836	0.4523

WM: Word translation model
PM_n: n-gram Phrase translation model

116

"Phrase" in IR

- Not necessarily consecutive words ("information …retrieval")
- Words at distance (context words) may be useful to determine the correct translation of a word

  - the new drug developed by X was approved by Y
  - drug … smuggle …

- Question
  - How to use context words (within a text window) to help translation?

  → Some approaches in SMT but will need to be further extended

117

Adding co-occurrences in query – one attempt

$$P(\mathbf{e}^D \mid Q) = \sum_{\mathbf{e}^Q \in Q} P(\mathbf{e}^D \mid \mathbf{e}^Q) P(\mathbf{e}^Q \mid Q)$$

$\mathbf{e}^Q$: term, bigram, proximity

$\mathbf{e}^D$: term

Query
instant messenger msn

instant messenger msn
instant-messenger messenger-msn
instant-messenger-msn
instant-msn

Title
msn web messenger

msn web messenger

118

Comparison of different translation models

	QE Models	NDCG@1	NDCG@3	NDCG@10
1	NoQE	0.2803	0.3430	0.4199
2	WM	0.3117	0.3717	0.4434
3	PM_2	0.3261	0.3832	0.4522
4	PM_8	0.3263	0.3836	0.4523
5	$CM_{T\text{-}B}$	0.3208	0.3786	0.4472
6	$CM_{T\text{-}P2}$	0.3204	0.3771	0.4469
7	$CM_{T\text{-}B\text{-}P3}$	0.3219	0.3790	0.4480
8	$CM_{T\text{-}B\text{-}P5}$	0.3271	0.3842	0.4534
9	$CM_{T\text{-}B\text{-}P8}$	0.3270	0.3844	0.4534

CM: Concept translation model (T-term, B-bigram, Pn-proximity within n)

119

Keeping phrases in translations

Query
instant messenger msn

instant messenger msn
instant-messenger messenger-msn
instant-messenger-msn
instant-msn

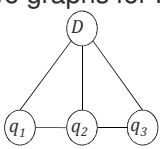
Title
msn web messenger

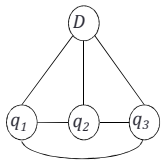
msn web messenger
msn-web web-messenger
msn-web-messenger
msn-messenger

120

Considering "phrases" in retrieval? - Markov Random Field

- Two graphs for IR





Sequential dependence Full dependence

$$Score(D, Q) = P_\Lambda(Q, D) = \frac{1}{Z} \sum_{c \in C(G)} \psi(c, \Lambda)$$

Markov Random Field

$$Score(D,Q) = P_\Lambda(Q,D) = \frac{1}{Z}\prod_{c\in C(G)} \psi(c;\Lambda)$$

$$\overset{rank}{=} \sum_{c\in C(G)} \lambda_c f(c) = \sum_{c\in T} \lambda_T f_T(c) + \sum_{c\in O} \lambda_O f_O(c) + \sum_{c\in U} \lambda_U f_U(c)$$

- Three components:
  - Term – T
  - Ordered term clique – O
  - Unordered term clique – U
- $\lambda_T, \lambda_O, \lambda_U$: fixed parameters
  - Typical good values (0.8, 0.1, 0.1)

$$f_T(c) = \log P(q_i \mid D)$$
$$f_O(c) = \log P(\#1(q_i,\ldots,q_{i+k}) \mid D)$$
$$f_U(c) = \log P(\#uw(q_i,\ldots,q_{i+k}) \mid D)$$

Results

	QE Models	NDCG@1	NDCG@3	NDCG@10
1	**MRF (NoQE)**	0.2802	0.3434	0.4201
2	**1 + M-CM$_{T\text{-}B\text{-}P8}$**	0.3293	0.3869	0.4548
3	**M-CM$_{T\text{-}B\text{-}P8}$**	0.3271	0.3843	0.4533

- Using MRF to evaluate "phrase" queries helps slightly.
- **1 + M-CM$_{T\text{-}B\text{-}P8}$**: Generate and use phrase translations
- **M-CM$_{T\text{-}B\text{-}P8}$**: Generate phrase translations but break them into bag of words

Discussions

- Monolingual IR can inspire from approaches to CLIR
  - Collecting "parallel" texts and train a translation model
- It is helpful to consider query and document being in 2 different languages
- Useful resources
  - Query logs (click-through)
  - If I don't have query logs?
  - Simulate queries by anchor texts (Croft et al.): anchor-page

General discussions on Web IR

- Multiple sources of information / knowledge
  - Documents
  - Hyperlinks / document structure
  - Query logs
    - Queries, query sessions, clickthrough
  - Users (communities)
- Dynamic nature of the Web
  - Up-to-date knowledge

→ See Talks on Web science and Web dynamics

125

Some applications of CLIR (1)

- CLIR in specialized areas
- Patent retrieval
  - Patents are written in local languages (Chinese, Korean, …)
  - Companies doing international business are supposed to be aware of the existing patents
  - CL patent retrieval in practical uses: increasing (patent searhc on Google)
- Medical IR
  - Relevant information in another language may be useful (Chinese medicine)
- Specific problems?
  - Patent/medical language?
  - Patent/medical document structure?
  - Term mismatch (standardization/expansion is important)

→Talk on domain-specific IR

126

Some applications of CLIR (2)

- CL Question-Answering
  - Yahoo! QA, Baidu knows, …
  - Some answers may only be available in a foreign language
  - Translate the question
    - Several possible formulations to select/combine

→ Talk on community QA

127

Some applications of CLIR (3)

- Image retrieval
- Mostly language-independent
- but the annotations and surrounding texts are language-dependent
- Query-Annotation search ~ CLIR problem
- Specific problem: annotations are limited
  → Expansion is required

128

Summary

- High-quality MT usually offers a good solution
- Well-trained TM based on parallel texts can match or outperform MT (Kraaij et al. 03)
- Dictionary
  - Simple utilization is not good
  - Translation selection is important → can match monolingual IR
- The performance of CLIR usually lower than monolingual IR (between 80% and 100% for advanced approaches)
- Better translation means → better CLIR effectiveness
- Better to combine translation means
- CLIR is now integrated in some search engines (wide use to be expected in the future)
- Remaining problems:
  - Improve translation quality (select good translation terms)
  - IR-oriented translation
- Beyond: CLIR is not separated from general IR

Some references

- Jian-Yun Nie, Cross-language information retrieval, Morgan-Claypool, 2010. (a survey book on CLIR)
- Adriani, M. van Rijsbergen, C.J., (2000). Phrase identification in cross-language information retrieval. In Proceedings of RIAO, pp. 520-528. (Using phrase in CLIR)
- Ballesteros, L. and Croft, W. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of SIGIR Conf. pp. 84-91. (Using phrases in CLIR)
- Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of SIGIR Conf.*, pp. 222-229. (Translation model for monolingual IR)
- Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), pp. 263-311. (IBM translation models)
- Chen, A., Jiang, H., and Gey, F. (2000). Combining multiple sources for short query translation in Chinese-English cross-language information retrieval. In Proceedings of the Fifth international Workshop on on information Retrieval with Asian Languages (IRAL). pp. 17-23 (Combining translation resources)
- Cui, H., Wen, J-R., Nie, J-Y, and Ma, W-Y. 2003. Query expansion by mining user log. *IEEE Trans on Knowledge and Data Engineering*. Vol. 15, No. 4. pp. 1-11. (Query expansion based on query logs)
- V. Dang and W.B. Croft. Query Reformulation Using Anchor Text. In Proc. of WSDM, pages 41-50, 2010. (simulating query logs by anchor texts)
- S. Dumais, T. Letsche, M. Littman, and T. Landauer, "Automatic cross-language retrieval using latent semantic indexing", in AAAI Symposium on Cross Language Text and Speech Retrieval. 1997, American Association for Artificial Intelligence. (Using LSI for CLIR)
- J. Gao, J-Y. Nie, E. Xun, J. Zhang, M. Zhou, C. Huang, Improving Query Translation for CLIR using Statistical Models, 24thACM-SIGIR, New Orleans, Sept. 2001, pp. 96-104. (selection of translation terms)
- Gao, J., He, X., and Nie, J-Y. 2010. Clickthrough-based translation models for web search: from word models to phrase models. In *CIKM*, pp. 1139-1148. (translation models for monolingual IR)
- E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In IJCAI (ESA)
- Gale, W. A., Church K. W. 1993. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(3): 75-102. (sentence alignment)
- Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks, In Proc. ASLIB translating and the computer 21 conference. (selection of translation terms)
- Jin, Rong, Hauptmann, A.G., and Zhai, CX. (2002). Title Language Model for Information Retrieval. In Proceedings of SIGIR Conf., pp. 42-48 (title language model)
- Kraaij, W., Nie, J.Y., and Simard, M. (2003). Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval. Computational Linguistics, 29(3): 381-420. (CLIR using language modeling and Web-mined parallel corpora)
- Liu, Y., Jin R. and Chai, Joyce Y. (2005). A maximum coference model for dictionary-based cross-language information retrieval, In Proceedings of SIGIR conf., pp. 536-543 (translation selection)
- J. Scott McCarley, Should we Translate the Documents or the Queries in Cross-language Information Retrieval? 1999, ACL, pp. 208-214. (Document translation vs. query translation)
- Nie, J.Y., Simard, M., Isabelle, P., Durand, R. (1999) Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web, In Proceedings of SIGIR Conf., pp. 74-81 (CLIR using Web-mined parallel pages)
- Pirkola, A. (1998) The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval." In Proceedings of SIGIR Conf., pp. 55-63. (Structured query translation)
- Pirkola, A., Toivonen, J., Keskustalo, H., Visala K., Järvelin K., (2003) Fuzzy translation of cross-lingual spelling variants, In Proceedings of SIGIR Conf., pp. 45-352. (Fuzzy match)
- Douglas W. Oard and Paul Hackett, "Document translation for the cross-language text retrieval at the university of Maryland", in the Sixth Text REtrieval Conference (TREC-6). (Document translation vs. Query translation)
- Riezler, S., and Liu, Y. 2010. Query rewriting using monolingual statistical machine translation. *Computational Linguistics*, 36(3): 569-582 (phrase translation for CLIR)
- Seo, H.-C., Kim, S.-B., Rim, H.-C. and Myaeng S.-H., (2005) Improving query translation in English-Korean cross-language information retrieval, Information Processing and Management, 41: 507-522. (translation selection)
- Sheridan, P. and Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In Proceedings of SIGIR Conf., pp. 58-65. (CLIR using comparable corpora)
- Valentin I. Spitkovsky and Angel X. Chang. 2012. A Cross-Lingual Dictionary for English Wikipedia Concepts. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012) (ESA for CLIR)
- Wen, J., Nie, J-Y., and Zhang, H. 2002. Query clustering using user logs. *ACM TOIS*, 20(1): 59-81. (Query clustering based on query logs)
- Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, and Robert E. Frederking, "Translingual information retrieval: Learning from bilingual corpora", Artificial Intelligence, vol.103, pp. 323–345, 1998. (Pseudo-RF for CLIR)