Cross-lingual and Multi-lingual IR

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

Definitions

- Cross-lingual (cross-language) IR (CLIR):
  Retrieval of documents in a language different from that of a query

  synonym: bilingual IR

- Multilingual IR (MLIR)
  Retrieval of documents in several languages from a query

outline

- Needs for CLIR and MLIR
- Problems in CLIR and MLIR
- Approaches to CLIR
  - MT
  - Bilingual dictionary
  - Parallel texts
- Approaches to MLIR
- Experiments and evaluation campaigns
- A better integration of translation and retrieval?

Needs for CLIR and MLIR

- Why is CLIR and MLIR useful?
  - An information searcher wants to retrieve relevant documents in whatever language.
  - Intelligence:
    - CIA,
    - companies (finding competing companies, finding calls for tenders, …)
  - A user speaking several languages also may want an MLIR to avoid typing the same query several times in different languages.

Problems in CLIR and MLIR

- CLIR and MLIR are based upon monolingual IR: all the problems of monolingual IR.
  Document representation v.s. query representation
- Problems due to the differences in languages.
  - Documents in E, F, I, …
  - Query in E

  1. Documents in F ⇓ document representation in E
     Query in E ⇓ query representation in E
  2. Query in E ⇓ query representation in F
     Documents in F ⇓ document representation in F

Key problem = translation

1. Document translation - Translate documents into the query language

Pros:

- translation may be (theoretically) more precise
- documents become "readable" by the user

Cons:

- huge volume to be translated
- impossible to translate them in all the languages (translate English documents in F, I, …, Chinese, Thai, …)

Key problem = translation (bis)

2. Query translation – Translate query into document language(s)

Pros:

- flexibility (translation on demand)
- less text to translate

Cons:

- less precise (2/3-word queries)
- The retrieved documents need to be translated (gist) to be readable.

How to translate

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

Approach 1: Using MT

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

  Query in F ⟶ Translation in E
  MT ↕
  Documents in E

- Problems:
  - Quality
  - Availability
  - Development cost

Word-by-word translation

- Select the first translation word
  - Assumption: The first translation is the most frequently used translation
  - Depends on the organization of the dict.
    - Not the case for Freedict:
      access: attaque, accéder, intelligence, entrée, accès
  - Problems:
    - May select a wrong translation
    - context-independent
    - May miss synomyms

Word/term level

- Choose one translation word
  - E.g. organic – organique
  - Better to keep all the synonyms (organique, biologique)? – query expansion effect
- Sometimes, use context to guide the selection of translation words
  - The boy grows: grandir
  - … grow potatoes: cultiver

Problems of MT

- Translation quality for CLIR and MLIR
  - Wrong choice of translation word/term
    - organic food – nouriture organique
    - train skilled personnel - personnel habile de train (ambiguity)
  - Wrong syntax
    - human-assisted machine translation - traduction automatique humain-aidée
  - Unknown words
    - Personal names:
      Bérégovoy ⇓ Bérégovoy, Beregovoy
      ⇓ Deng Xiaoping, Deng Hsao-ping, Deng Hsiao p'ing

Word-by-word translation (bis)

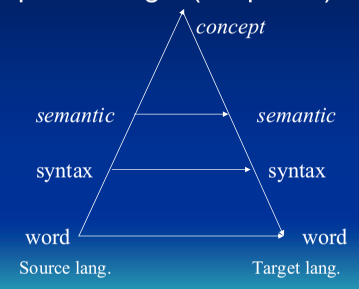
- Concatenate all the translation words
  access: attaque, accéder, intelligence, entrée, accès
  - Covers all the possible translation
  - Keeps ambiguity and incorrect translations
  - Noisy query translation

Syntax – unused effort for CLIR?

- Current IR approaches based on words (bag of words)
- Efforts on determining the correct syntax not used for IR
- However, useful for disambiguation (stem international terrorism v.s. tree stem)

State of the art of MT

- Vauquois triangle (simplified)



Translate the query as a whole

Best global translation for the whole query

- Candidates:
  For each query word
  - Determine all the possible translations (through a dict.)

2. Selection
   select the set of translation words that produce the highest ***cohesion***

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
- Approaches
  - For each word in a query
    - Select the first translation word
    - Select all the translation words
  - For all the query words
    - Select the translation words that create the highest cohesion

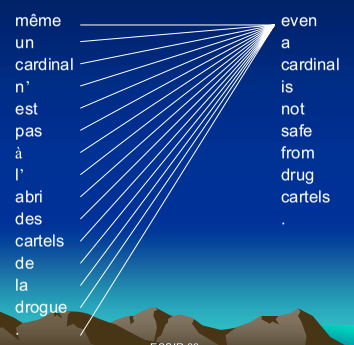
State of the art of MT (cont'd)

- General approach:
  - Word / term: dictionary
  - Syntagm (phrase)
  - Syntax
  - "semantic"

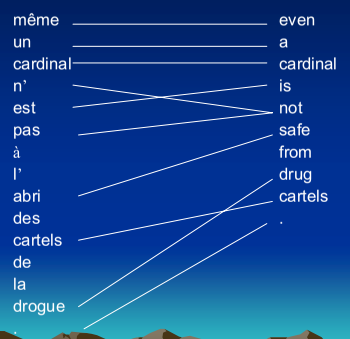
Example of aligned sentences

Débat L'intelligence artificielle	Artificial intelligence A Debat
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.
	Two further points are important.
Les symboles et les programmes sont des notions purement abstraites.	First, symbols and programs are purely abstract notions.

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



Application of EM: $p(t_j|s_i, A)$



Parallel texts (cont'd)

- Training a translation model
- Principle:
  - train a statistical translation model from a set of parallel texts: $p(t_j|s_i)$
  - Principle: The more s_j appears in parallel texts of t_i, the higher $p(t_j|s_i)$.
- Given a query, use the translation words with the highest probabilities as its translation

Principle of model training

- $p(t_j|s_i)$ 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: $S_k \longleftrightarrow T_l$
  - Initial probability assignment: $t(t_j|s_i, A)$
  - Expectation Maximization (EM): $p(t_j|s_i, A)$
  - Final result: $p(t_j|s_i) = p(t_j|s_i, A)$

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

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)

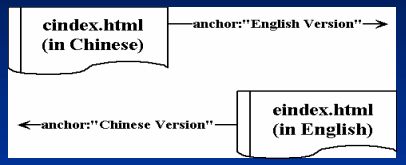
Approach 3: parallel texts

- Parallel texts contain possible translations of query words
- First exploration: using IR methods
  - 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

parallel texts (cont'd)

- Second approach: LSI
  - In monolingual LSI, singular value decomposition (SVD) is able to group synonyms in the created structure
  - For CLIR, SVD is performed on a parallel corpus
    - The LSI encompasses translation relationships (special case of cross-language synonym)
  - Query in F can match directly documents in E in LSI
- Problems:
  - Computational complexity
  - Number of singular value to choose (empirical setting)
  - Coverage of the parallel texts w.r.t. semantics

Candidate Sites Searching



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

An example of "parallel" pages

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

IBM models

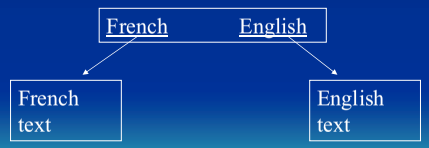
- IBM 1: does not consider positional information and sentence length
- IBM 2: considers sentence length and word position
- IBM 3, 4, 5: fertility in translation
- For CLIR, IBM 1 seems to correspond to the current approaches of IR.

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

STRANDS

- Assumption:
  If - A Web page contains 2 pointers
  - The anchor text of each pointer identifies a language
  Then The two pages referenced are "parallel"



How effective is this approach? (With the Hansard)

	F-E (Trec6)	F-E (Trec7)	E-F (trec6)	E-F (Trec7)
Monolingual	0.2865	0.3202	0.3686	0.2764
Dict.	0.1707 (59.0%)	0.1701 (53.1%)	0.2305 (62.5%)	0.1352 (48.9%)
Systran	0.3098 (107.0%)	0.3293 (102.8)	0.2727 (74.0%)	0.2327 (84.2%)
Hansard TM	0.2166 (74.8%)	0.3124 (97.6%)	0.2501 (67.9%)	0.2587 (93.6%)
Hansard TM+ dict.	0.2560 (88.4%)	0.3245 (101.3%)	0.3053 (82.8%)	0.2649 (95.8%)

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

PTMiner

- 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.

Problem of parallel texts

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

Summary of the experimental results

- High-quality MT is still the best solution
- TM based on parallel texts can match MT
- Dictionary
  - Simple utilization is not good
  - Complex approaches improve quality
- The performance of CLIR usually lower than monolingual IR (between 50% and 90% of monolingual in general)

CLIR Results: C-E

- Test collections from TREC
  - Chinese: People's Daily, Xinhua news agency
  - English: AP

	C-E	E-C
Monolingual	0.3861	0.3976
Dictionary (EDict)	0.1530 (39.6%)	0.1427 (35.9%)
TM	0.2063 (53.4%)	0.2013 (50.6%)
TM + Dict	0.2811 (72.8%)	0.2601 (65.4%)

  - MT system:
    E-C: 0.2001 (50.3%) C-E: (56 - 70%)

Further verification of parallelism

- Download files (for verification with document contents)
- Compare file lengths
- Check file languages (by an automatic language detector – SILC)
- Compare HTML structures
- (Sentence alignment)

Summary of the existing approaches to CLIR

CLIR = Query Translation + IR

- Trend: Integrate QT with IR
  - QT is one step in the global IR process
  - E.g. Kraaij, Nie and Simard, 2003
  - Using language model

Problems of using parallel corpora

- Not strictly parallel (Web)
- Coverage
- In a different domain than the documents to be retrieved
- Not applicable to "minor" languages

Mining Results

- 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, …

CLIR as a special case of query expansion

- Query expansion:
  - Q ⟶ Q' ⟷ D
    related terms
- CLIR
  - Q ⟶ Q' ⟷ D
    trans. rel.
  - Translation relation ~ term relation
- Inferential IR
  - Infer an expression Q' such that when Q' is satisfied, Q is too.

Translation problems with TM

- Compound terms (e.g. *pomme de terre* – earth, apple, potato, soil, …)
- Coverage (personal names, unknown words)
- Ambiguous words remain ambiguous (*drug* – médicament, drogue)
- Possible solutions
  - Recognize compounds before model training
  - treat named entities differently
  - Combine with dictionaries

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
Systran	0.3098 (107.0%)	0.3293 (102.8)	0.2727 (74.0%)	0.2327 (84.2%)
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

Future problems

- Develop better translation tools for IR (e.g. for special types of data such as personal names)
- Integrating multiple translation results
- Translate non-English languages
- Integration of query translation and retrieval process
- Develop approaches to MLIR
- Make the retrieved documents readable

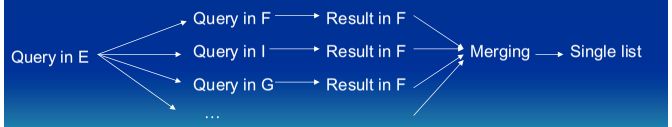
Experiments

- CORI works well for monolingual distributed IR
- For MLIR, Raw score and Normalized score work better than CORI in CLEF01 and CLEF02
- The effectiveness of MLIR is lower than CLIR (bilingual IR)

Multilingual IR

MLIR = CLIR + merging

- Translate the query into different languages
- Retrieve doc. in each language
- Merge the results into a single list



Some References

- Introduction and survey:
  - D. Oard, several survey papers http://www.glue.umd.edu/~oard/research.html#overviews
- MT-based approaches
  - (Chen 02) A. Chen, Cross-language retrieval experiments at CLEF 2002, in CLEF-2002 working notes, pp. 5-20, 2002.
  - (Savoy 02), J. Savoy, Report on CLEF-2002 experiments: Combining multiple sources of evidence, in CLEF-2002 working notes, pp. 31-46., 2002
- Dictionary-based approaches
  - (Grefenstette 99) G. Grefenstette, The WWW as a resource for example-based MT tasks, proc. ASLIB translating and the computer 21 conference, London, 1999.
  - (Qu et al. 02) Y. Qu, G. Grefenstette, D. Evens, Resolving translation ambiguity using monolingual corpora – A report on Clairvoyance CLEF-2002 experiments, in CLEF-2002 working notes, 2002, pp. 115 – 126.
  - (Gao et al. 02) J. Gao, J.-Y., Nie, H. He, W. Chen, M. Zhou,Resolving Query Translation Ambiguity using a Decaying Co-occurrence Model and Syntactic Dependence Relations, 25th ACM-SIGIR, 2002, Tampere, pp. 183-190.
- Based on parallel texts
  - (Nie et al. 99) J.Y. Nie, M. Simard, P. Isabelle, R. Durand, "Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web", 22ndACM-SIGIR, Berkeley, 1999, pp. 74-81
  - (Kraaij et al 03) W. Kraaij, J.-Y. Nie, M. Simard, Embedding Web-based Statistical Translation Models in Cross-Language Information Retrieval, Computational Linguistics, 29(3), 2003.
  - (Yang 98) Y. Yang, J.G. Carbonell, R.D. Brown, R.E. Frederking, Translingual information retrieval: learning from bilingual corpora, *Artificial Intelligence*, 103: 323-345, 1998.
  - (Nie 02) J.-Y. Nie, F. Jin, A multilingual approach to multilingual information retrieval, *Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*, Rome, Sept. 2002, pp. 101-110.

MLIR (cont'd)

- MLIR = mixed query for mixed doc. Collection (Chen 02, Nie 02)
  - Translate the query into all the languages
  - Concatenate them into a mixed query
  - IR using mixed query on mixed documents

  - Avoiding merging
  - homograph in different languages (but, pour, …)
  - Possible improvement: distinguishing language (add a tag to the indexes, e.g. but_f, pour_e)

Merging – often used approaches

- Round-robin
  - Take the first from the list of F, E, I, …
  - Take the second from the list of F, E, I, …
  - …

  Assumption: similar number of rel. doc., ranked similarly
- Raw score
  - Mix all the lists together
  - Sort according to the similarity score

  Assumption: similar IR method, collection statistics

- LSI
  - M. Littman and S. Dumais and T. Landauer, Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing, in G. Gredenstette (Ed.), Cross Language Information Retrieval, 1997, http://citeseer.nj.nec.com/dumais97automatic.html
  - Mori et al. CLIR based on LSI with multiple space, NTCIR -2, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/mori-ir.pdf
- CLIR & MLIR Campaigns
  - TREC CLIR track (http://trec.nist.gov)
  - CLEF (http://videoserver.iei.pi.cnr.it:2002/DELOS/CLEF/)
  - NTCIR (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/)
- Sentence alignment
  - (Gale & Church, 93) W. A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, Computational Linguistics, 19 :1, 75-102, 1993.
  - (Simard et al.92) M. Simard, G. Foster, P. Isabelle, Using Cognates to Align Sentences in Parallel Corpora, Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, 1992.
- Translation models
  - (Brown et al, 93) P. F. Brown, S.A.D. Pietra, V. D. J. Pietra, and R. L. Mercer, The mathematics of machine translation: Parameter estimation. Computational Linguistics, vol. 19, pp. 263-312, 1992.
- Parallel text mining
  - P. Resnik, Parallel Stands: A preliminary investigation into mining the Web for bilingual text, AMTA'98, 1998.
  - J. Chen, J.Y. Nie, Web parallel text mining for Chinese-English cross-language information retrieval, NAACL-ANLP, Seattle, May 2000.

Problems in MLIR

- Translation
- Merging different (incompatible) retrieval results
  - Is it necessary to produce a single mixed result list ?

Merging (cont'd)

- Normalized score
  - S' = S / S_max
  - S' = (S – S_min) / (S_max – S_min)
- CORI
  - S' = S * (1 + (S – S_avg) / S_avg)

  Idea: modify the raw score according to the average score for a collection (language)