Cross Language Information Retrieval

Raffaella Bernardi

Università degli Studi di Trento

P.zza Venezia, Room: 2.05, e-mail: bernardi@disi.unitn.it

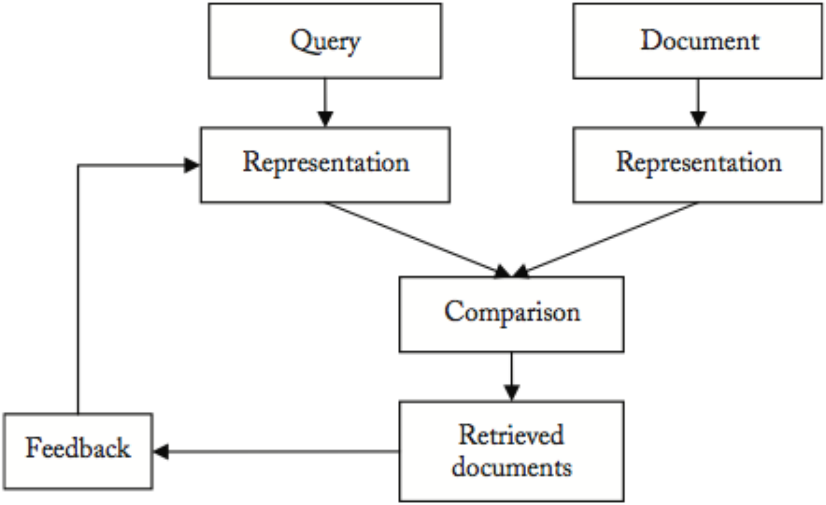
Contents

1. Acknowledgment

The slides of this lecture summarize some of the chapters in:

> Jian-Yun Nie "Cross-Language Information Retrieval". Synthesis Lectures in Human Language Technologies nr. 8. Morgan & Claypool Publishers.

2. IR in a picture



3. Cross Language IR

Queries and documents are described in different languages. E.g.:

> query: “major earthquakes in recent year”

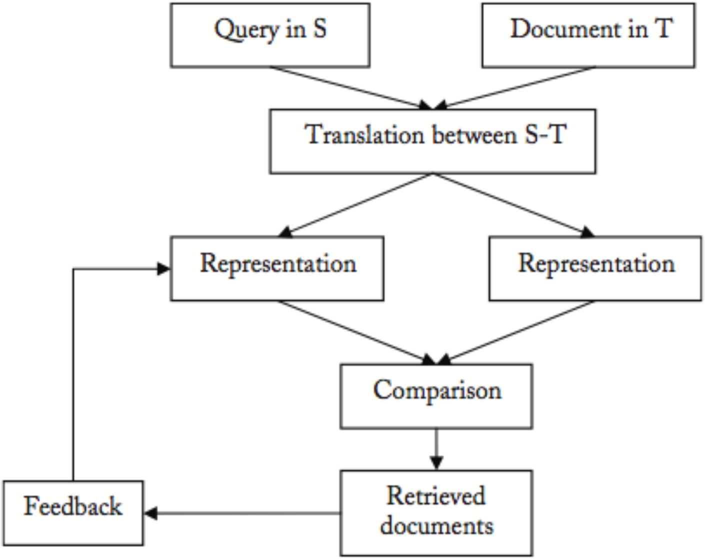
we want to retrieve both passages below:

> There is a major earthquake in Wenchuan, China in 2008 (EN)
>
> Un tremblement de terre violent à Wenchuan secoue la Chine en 2008 (FR)

3.1. Motivations

- searching for tagged images.
- the relevant document may not exist in the user's native language
- user intend to find all the relevant information available, whatever language is used (e.g. searching for patents)
- user could be able to read documents in language different from the native one, but could have difficulty in formulating queries in those languages.

3.2. CLIR in a picture



3.3. The translation module

The translation module could use one of the following approaches:

Query translation Mapping the query representation into the document representation

- Pro: flexible, more interaction with the user (who could choose the languages of interest, can correct the translation.)
- Contra: translation ambiguity amplified by the lack of context.

Document translation Mapping the document representation into the query rep.

- Pro: more context but current MT systems don't exploit the context much
- Contra: one has to determine in advance to which language each document should be translated, all the translated versions should be stored.

Inter-lingua translation Mapping document and query representation to a 3rd language

- Pro: useful if there is no resource for a direct translation.
- Contra: lower performance that the direct translation

Most used approach in CLIR: Query translation

3.4. CLIR History

1. Started in 1960s, CLIR research has been particular active in the area of library science.

2. Took off from mid-1990s with the World-Wide Web.

3. 1997 started CLIR at TREC organized by the National Institute of Standards and Technology (NIST)

4. 2000 CLEF (Cross-Language Experiment Forum) CLIR experiments on European languages: first experiment: EN, GE, IT, SP, Swedish and Finish. Then more and more languages. In CLEF 2007: Indian languages too.

5. 1999 NTCIR: Asian languages

CLIR effectiveness (MAP) was much lower than that of monolingual IR (around 50%); nowadays they are close (80%-100%).

3.5. Companies & CLIR

- 2006 Yahoo!: translates queries in French and German to English, Spanish, Italian and French/German
- 2007 Google: searching without boundaries
- here: CLS (Cross Library Service)

3.6. Query Translation: Solutions

- Machine Translation
- Dictionary-based translation
- Translation based on parallel or comparable corpora

4. Query translation: MT difficulties

Translation in CLIR is not a traditional translation task:

- No much context
- High ambiguity
- Desired query expansion effect (no unique translation)
- Term weighting

Hence, Machine Translation (MT) tools are not appropriate for query translation.

5. Query Translation: Dictionary-based

For CLIR, dictionaries are usually considered as word list. E.g.

FR	EN
accent:	accent, stress
accentuer	accent, accentuate, stress
accepter:	accept, receive, take, take in

a) Using all the translations for each query word;

b) Using the first translation listed in the dictionary.

a) corresponds to expanding the query, but also including its different meanings. Hence, increase in recall at the cost of decreasing precision. Not effective strategy.

b) the first translation is usually the most frequently used. (but the order depends on how the dictionary is organized.)

Both methods are insufficient.

5.1. Term weighting problem

The fact that more translations are included for a term has an impact on the relative importance of the term in the query even if that was not intended in the original query. E.g.

“problem of AIDS prevention” the translation will contain far more translation words for “problem” than for “AIDS”. So the translation of “problem” will dominate those of “AIDS”.

This problem could be solved by normalizing the term weights for translations per source term (weight: 1/n where n is the nr of translation for the source term.) [Xu and Weischedel 2005]

5.2. Coverage of the Dictionary problem

Xu and Weischedel (2005) have shown that: once the dictionary reaches certain coverage of frequent source-language words, the further the increase of the size of the dictionary will not improve more the CLIR effectiveness. (tried with English-Chinese CLIR and English-Arabic (ca 10K entries).) NOTE: The test queries are those of TREC that uses quite frequent English terms.

- specialized terms may be missing
- translation stored could still be ambiguous
- phrase may not be translated correctly

5.3. Ambiguity problem

Example:

> “chamber” (FR) most frequent EN tr. “bedroom”, but “musique de chambre” is not “bedroom music”, similarly for “chambdre de commerce”.

Grefenstette (1999): the correct translation words usually have a higher frequency of co-occurrences with other translation words. Similarly, use similarity measures.

Others: for longer query, compute the similarity only for e.g. words within a noun phrases (parsed query.) or words in a dependency relation.

6. Query Translation: parallel and comparable corpora

Parallel Corpora are texts with their translations in another language and with translation relations between texts, sentences, phrases and words.

Comparable Corpora are texts that are topically similar without being parallel.

Parallel corpora have been widely exploited in MT since 1990s. Translation usually uses the same paragraph and sentence order.

1. Algorithms are developed to align them.
2. MT models are trained on the aligned corpora.

6.1. Translation models for CLIR trained on parallel corpora

IBM model : no word order is considered, hence a word that appears in a sentence can be translated by any of the words in the aligned sentences in other language.

IBM model is the most used in CLIR : by allowing looser translation, we extend the strict translation to word that co-occur in the aligned sentence.

Drawback of IBM model for CLIR : common words in the target language receive high translation probability for many source language words.

Solution : supplement the statistically based translation model (larger coverage) with a bilingual dictionary (more precise translations), to overcome the ambiguity problem similarity measures are used.)

Sum up The translation probabilities are considered as weights of terms, which are used in the subsequent monolingual IR process.

6.2. Comparable corpora

- Aim: determine cross-language similarity between terms in two languages.
- The more two terms co-occur in the comparable texts simultaneously, the more they are assumed to be similar in meaning.
- Compute co-occurrences for a pair of terms in two different languages.
- each query term is replaced by the set of (weighted) similar terms in another language.

Quite noisy results. They are used only if no parallel corpus is available.

New line: Mining the web to build parallel corpora.

7. Pre- and post-translation expansion

Pre-translation expansion it uses the original query to retrieve a set of documents from a collection in the source language. A set of terms are extracted from them and added into the query before translation.

Post-translation expansion a set of target-language documents are retrieved using the translated query, and a set of terms are extracted form them and used to enhance the query. A new retrieval is finally performed to retrieve a new set of documents.

Both methods have been shown to improve both precision and recall.

8. Evaluation: TREC

Query Translation and IR: In general, for resource-rich language pairs, CLIR systems have achieved retrieval effectiveness at the level of 80-100% of that of monolingual IR.

Document Translation: aiming at making the retrieved documents understandable to the user. A lot to be improved.

Future: aiming at developing translation models specifically for CLIR and well integrated in it. Query translation is similar to query (monolingual) expansion.

9. Evaluation: translation service in search

[Savoy and Dolamic 2009] "How effective is Google's translation service in search?"

- French newspaper Le Monde, Swiss news (1994-1995). 177,452 documents. Av. 178 content-bearing term per doc.
- 299 queries
- Relevance judgments by humans.
- MRR (Mean reciprocal rank): this value served as a measure of any given search engines ability to extract one correct answer and list it among the top-ranked items. For any given query, r is the rank of the first relevant document retrieved and the query performance is computes as $1/r$.
- MAP (Mean average precision): it accounts for the rank of all relevant items.

9.1. Google: EN to FR

	MRR		MAP	
	Monolingual FR	From EN queries	Monolingual FR	From EN queries
Okapi	0.66	0.58	0.40	0.34
Language Model	0.59	0.50	0.36	0.30
tf-idf	0.50	0.38	0.25	0.20

Note: the two languages have similar meanings and similar spelling. Proper names have comparable spellings. Typical mistakes: translate (e.g. “European Cup”, “coffee cup” same translation regardless of the context; Turkey (country) vs. turkey (animal): case sensitive, but not always ok. idioms and compound terms)

10. Administrativa

- Wednesday 9th: There is the CdF in the pm. 16:30-18:00? (or in the morning 09:00-10:30?)
- Wednesday 16th: There is the last lecture. Morning (10:00-12:00)? Or pm (16:00-18:00?)