Introduction to Information Retrieval

http://informationretrieval.org

Cross Language IR

Hinrich Schütze, Christina Lioma

Institute for Natural Language Processing, University of Stuttgart

2010-07-05

Outline

Definitions

- Crosslingual (a.k.a. cross-language) IR (CLIR)
- Multilingual (a.k.a. multi-language) IR (MLIR)

Definitions

- Crosslingual (a.k.a. cross-language) IR (CLIR): retrieval of documents in a language different from that of a query. E.g., bilingual or trilingual IR
- Multilingual (a.k.a. multi-language) IR (MLIR)

Definitions

- Crosslingual (a.k.a. cross-language) IR (CLIR):
  retrieval of documents in a language different from that of a query. E.g., bilingual or trilingual IR
- Multilingual (a.k.a. multi-language) IR (MLIR):
  retrieval of documents in several languages

Definitions

- Crosslingual (a.k.a. cross-language) IR (CLIR): retrieval of documents in a language different from that of a query. E.g., bilingual or trilingual IR
- Multilingual (a.k.a. multi-language) IR (MLIR): retrieval of documents in several languages

Motivation

Internet usage: 29.5% English, 70.5% non-English (Lazarinis et al. 2007)

Definitions

- Crosslingual (a.k.a. cross-language) IR (CLIR): retrieval of documents in a language different from that of a query. E.g., bilingual or trilingual IR
- Multilingual (a.k.a. multi-language) IR (MLIR): retrieval of documents in several languages

Motivation

Internet usage: 29.5% English, 70.5% non-English (Lazarinis et al. 2007)

- user scenarios: monolingual / multilingual users (partly or passively)

Definitions

- Crosslingual (a.k.a. cross-language) IR (CLIR): retrieval of documents in a language different from that of a query. E.g., bilingual or trilingual IR
- Multilingual (a.k.a. multi-language) IR (MLIR): retrieval of documents in several languages

Motivation

Internet usage: 29.5% English, 70.5% non-English (Lazarinis et al. 2007)

- user scenarios: monolingual / multilingual users (partly or passively)
- intelligence:
  - state
  - companies (finding competing companies, finding calls for tenders, etc...)

Outline

Language-specific problems

1. Encoding, capitalisation, diacritics, compounding, complicated morphology ...

Language-specific problems

1. Encoding, capitalisation, diacritics, compounding, complicated morphology ...
2. Lack of script standards, e.g. `Khrushchev`, `Chrustschev`, `Khrooshtchoff`, `Chruhszhtchow`, `Jruchev`, `Chroesjtjov`, `Crustsciof`

Language-specific problems

1. Encoding, capitalisation, diacritics, compounding, complicated morphology ...
2. Lack of script standards, e.g. `Khrushchev`, `Chrustschev`, `Khrooshtchoff`, `Chruhszhtchow`, `Jruchev`, `Chroesjtjov`, `Crustsciof`
   - **Transliteration:** spelling words from one language with characters from the alphabet of another, usually in a character-by-character replacement

Language-specific problems

1. Encoding, capitalisation, diacritics, compounding, complicated morphology ...
2. Lack of script standards, e.g. `Khrushchev`, `Chrustschev`, `Khrooshtchoff`, `Chruhszhtchow`, `Jruchev`, `Chroesjtjov`, `Crustsciof`
   - **Transliteration:** spelling words from one language with characters from the alphabet of another, usually in a character-by-character replacement
   - **Transcription:** representation of the sound of words in a language using any set of symbols, i.e., the International Phonetic Alphabet (IPA)

Language-specific problems

1. Encoding, capitalisation, diacritics, compounding, complicated morphology ...
2. Lack of script standards, e.g. `Khrushchev`, `Chrustschev`, `Khrooshtchoff`, `Chruhszhtchow`, `Jruchev`, `Chroesjtjov`, `Crustsciof`
   - **Transliteration:** spelling words from one language with characters from the alphabet of another, usually in a character-by-character replacement
   - **Transcription:** representation of the sound of words in a language using any set of symbols, i.e., the International Phonetic Alphabet (IPA)
   - Latin script predominance on the Web, e.g. Greeklish
   - Often adhoc use of numbers and symbols, e.g. 8 for θ

Language-specific problems

3 Not always one-to-one correspondence with Latin characters, e.g., standard Hebrew (undotted & unvocalised) orthography

Language-specific problems

3. Not always one-to-one correspondence with Latin characters, e.g., standard Hebrew (undotted & unvocalised) orthography
4. Writing order:
   - Standard Indo-European: top-to-bottom, left-to-right
   - Hebrew, Japanese: right-to-left

Language-specific problems

3. Not always one-to-one correspondence with Latin characters, e.g., standard Hebrew (undotted & unvocalised) orthography
4. Writing order:
   - Standard Indo-European: top-to-bottom, left-to-right
   - Hebrew, Japanese: right-to-left
5. Need tokenisation
   - Arabic, Iranian, Uzbeki (use variants of the Arabic script): no capitalisation, no punctuation, hence difficult to detect sentence boundaries. Also, letters may be joined: letter looks different when it stands alone, when it is the first letter of a connected set of letters, when it is somewhere in the middle of a connection, and when it appears at the end of a set of connected letters.
   - costly, may introduce error

Language-specific problems

6 Under-represented languages

Language-specific problems

6 Under-represented languages

Example

Armenian uses its own script (its own I-E branch): not widely known in the world

- Small number of native speakers (3 million in Armenia, 8 million abroad)
- Changes in the script: 1920s Soviet Armenia reformed spelling, which however was rejected by the Armenian diaspora (which outnumbers significantly the country's population)

Result: already weak presence of Armenian on the Web lacks uniformity in script, which practically means noise for search engines.

Outline

IR problems

IR problems arising from non-standard script

- The same language entities are represented under different forms: no new words are added to the language, only different ways of writing the same words

IR problems

IR problems arising from non-standard script

- The same language entities are represented under different forms: no new words are added to the language, only different ways of writing the same words
- **Indexing problem:** Should all these term variants be indexed as one entry or as separate entries? Should these terms be normalised in some way, e.g., stemmed?

IR problems

IR problems arising from non-standard script

- The same language entities are represented under different forms: no new words are added to the language, only different ways of writing the same words
- **Indexing problem:** Should all these term variants be indexed as one entry or as separate entries? Should these terms be normalised in some way, e.g., stemmed?
- **Matching problem:** Should a query containing the term in Russian letters be matched to a relevant document containing the term in Latin letters? Should a term written in Russian letters receive the same term weight as the same term written in Latin letters?

Solution: key problem = translation

Treat as monolingual IR with translation

Solution: key problem = translation

Treat as monolingual IR with translation

1. Document translation - translate documents into the query language

Solution: key problem = translation

Treat as monolingual IR with translation

1. Document translation - translate documents into the query language

- Advantages:
  - Translation may be more precise (in principle)
  - Documents become readable by the user
- Disadvantages:
  - Huge volume to be translated
  - Impossible to translate them in all languages (Eng $\rightarrow$ Fre, Ger, Ita...)

Solution: key problem = translation

2. Query translation - translate query into the document language(s)

Solution: key problem = translation

2. Query translation - translate query into the document language(s)

- Advantages:
  - Flexibility (translation on demand)
  - Less text to translate
- Disadvantages:
  - Less precise (2-3-word queries)
  - The retrieved documents need to be translated (gist) to be readable

Integration of translation to IR

Approach 1:

- translate the query into different languages
- retrieve doc. in each language
- merge the results into a single file

Integration of translation to IR

Approach 1:

- translate the query into different languages
- retrieve doc. in each language
- merge the results into a single file
- round-robin: take the first from each list, then the second, and so on... Assumption: similar number of documents ranked similarly
- raw score: mix all the lists together and sort according to the similarity score. Assumption: similar IR method & collection statistics

Integration of translation to IR

Approach 2:

- translate the query into all the languages
- concatenate them into a mixed query
- IR using mixed query on mixed documents

Integration of translation to IR

Approach 2:

- translate the query into all the languages
- concatenate them into a mixed query
- IR using mixed query on mixed documents
- avoid merging
- homograph in different languages (`but`, `pour`, ...)
- possible improvement: distinguish language (e.g. add a tag to the index, e.g. `but_f`, `pour_e`)

Outline

How to translate

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

Approach 1: using MT

- Good solution iff translation quality is high

Approach 1: using MT

- Good solution iff translation quality is high
- Problems:
  - Quality
  - Availability
  - Development cost

Problems of MT

- Translation quality

Problems of MT

- Translation quality
  - Wrong choice of translation word/term
    - `organic food` → `nourriture organique` ambiguity

Problems of MT

- Translation quality
  - Wrong choice of translation word/term
    - `organic food` → `nourriture organique` ambiguity
  - Wrong syntax
    - `Human-assisted machine translation` → `traduction automatique humain-aideé`

Problems of MT

- Translation quality
  - Wrong choice of translation word/term
    - `organic food` → `nourriture organique` ambiguity
  - Wrong syntax
    - `Human-assisted machine translation` → `traduction automatique humain-aideé`
  - Unknown words
    - Personal names
    - Transliteration, transcription

Approach 2: using bilingual dictionaries

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

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

Approach 2: using bilingual dictionaries

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

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

- Approaches
  - for each word in a query
    1. select the best translation word
    2. select all the translation words

Approach 2: using bilingual dictionaries

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

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

- Approaches
  - for each word in a query
    1. select the best translation word
    2. select all the translation words
  - for all query words
    - select the translation words that create the highest cohesion

Cohesion

cohesion $\sim$ frequency of two translation words together

Example

- `data: données, matériau, data`
- `access: attaque, accéder, entrée, accès`

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

...

Approach 3: parallel texts

Parallel texts contain possible translations of query words

Approach 3: parallel texts

Parallel texts contain possible translations of query words

- Given a query in French
- Find relevant documents in the parallel corpus
- Extract keywords from their parallel documents, and consider them as a query translation

Parallel texts (cont.)

Training a translation model

- Principle:
  - Train a statistical translation model from a set of parallel texts: $p(t_j|s_i)$
  - The more s_i appears in parallel texts of t_j, 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:

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 = parallel texts
  - Sentence alignment A: $S_k \leftrightarrow T_h$
  - Initial probability assignment: $t(t_j|s_i, A)$
  - Expectation Maximisation (EM): $p(t_j|s_i, A)$
  - Final result: $p(t_j|s_i) = p(t_j|s_i, A)$

Sentence alignment

Assumptions:

1. The order of sentences in two parallel texts is similar
2. A sentence and its translation have similar length (length-based alignment)
3. A translation contains some 'known' translation words or cognates

Effectiveness: mean average precision

	F-E (TREC6)	F-E (TREC7)	E-F (TREC6)	E-F (TREC7)
monolingual	0.2865	0.3203	0.3686	0.2764
Dict.	0.1707	0.1701	0.2305	0.1352
Systran	0.3098	0.3293	0.2727	0.2327
Hansard PT	0.2166	0.3124	0.2501	0.2587
Hansard PT+dict	0.2560	0.3245	0.3053	0.2649

Problem of parallel texts

- Only a few large parallel corpora (e.g. Canadian Hansards, EU parliament, HK Hansards, UN documents...)
- Minor languages are not covered

Problem of parallel texts

- Only a few large parallel corpora (e.g. Canadian Hansards, EU parliament, HK Hansards, UN documents...)
- Minor languages are not covered
- Is it possible to extract parallel texts from the WEB?
  - STRANDS: If a Web page contains two pointers, the anchor text of each pointer identifies a language. Then, the two pages are references as parallel
  - PTMiner: parallel web pages = similar URLs at the difference of a tag identifying a language
    - `index.html` vs. `index_f.html`
    - `/english/index.html` vs. `/french/index.html`

Mining results (Nie 2003)

- French - English
  - Exploration of 30% of 5474 candidate sites
  - 14198 pairs of parallel pages
  - 135MB French texts and 118MB English texts
- Chinese - English
  - 196 candidate sites
  - 14820 pairs of parallel pages
  - 117.2M Chinese texts and 136.5M English texts

CLIR results: F-E

	F-E (TREC6)	F-E (TREC7)	E-F (TREC6)	E-F (TREC7)
monolingual	0.2865	0.3203	0.3686	0.2764
Dict.	0.1707	0.1701	0.2305	0.1352
Systran	0.3098	0.3293	0.2727	0.2327
Hansard PT	0.2166	0.3124	0.2501	0.2587
Web PT	0.2389	0.3146	0.2504	0.2289

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

Summary

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

Wrap up

- 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