Information Retrieval (IR)

- "Academic discipline that researches models and methods to access and organize large amounts of unstructured and structured information"
- Access is by using queries (these are a more or less appropriate statements of user's information need)
- Issues:
  - mismatch between document and query due to language ambiguity (synonym, homograph, homonym, paraphrasing, typo)
  - mismatch between document and query due to incomplete understanding of problem ("garbage in, garbage out")
  - Noisy document collection (OCR)
  - misleading content (spam etc.)
  - authority, source, actuality, copyright
  - relevance is subjective and context-dependent
The MLIR Challenge

"Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified."


MLIR / CLIR

- Monolingual retrieval in non-English languages
- Bilingual retrieval A → B
- Multilingual retrieval A → A, B, ...
- Multilingual retrieval AB → A, AB, AC, B, BC, ...
- Multilingual Information Access / Multilingual Retrieval encompasses all four definitions
- Cross-Language Information Retrieval (CLIR) means at least a bilingual retrieval between two different languages
- We can translate: queries, documents, both, neither!
- The "simplest scenario" translate the query (QT)

One Possible MLIR Flow

MLIR Reality

- Strč prst skrz krk
- Mitä sinä teet?
- Mam swoją książkę
- Nem fáj a fogad?
- Er du ikke en riktig nordmann?
- Добре дошли в България!
- Fortuna caeca est
- 我不是中国人
MLIR Reality

- Bilingual / multilingual (europa.eu/abc/
- Many countries are bi-/ multilingual (Canada (2), Singapore (2), India (21), EU (23))
- Official languages in EU: Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, and Swedish.
- Other languages: Catalan, Galician, Basque, Welsh, Scottish, Gaelic, Russian.
- Working languages in EU (mainly): English, German, French;
- In UN: Arabic, Chinese, English, French, Russian, Spanish.
- Court decisions written in different languages
- Organizations: FIFA, WTO, Nestlé, …

MLIR Reality

- Cases of multilingual IR
  - people may express their needs in one language and understand another
  - we may write a query in one language and understand answer given in another (e.g., very short text in QA, summary statistics, factual information (e.g., travel))
  - There are language-independent media that may be described in a different language (image, video, music)
  - to have a general idea about the contents (and latter to manually translate the most pertinent documents)
  - more important with the Web (however consumers prefer having the information in their own language).

Evaluation Campaigns

- TREC (trec.nist.gov)
  - TREC 3-5: Spanish
  - TREC 5-6: Chinese (simplified, GB)
  - TREC 6-8: Cross-lingual (EN, DE, FR, IT)
  - TREC 9: Chinese (traditional, BIG5)
  - TREC 10-11: Arabic
  - See [Harman 2005]

- Objectives
  - Promote IR research & communication with industry
  - Speed the transfer of technology
  - Build larger test-collections (evaluation methodology)

Internet users by language (www.internetworldstat.com)
Evaluation Campaigns

- **CLEF** (www.clef-initiative.eu)
  - Started in 2000 with EN, DE, FR, IT
  - 2001-02: EN, DE, FR, IT, SP, NL, FI, SW
  - 2003: DE, FR, IT, SP, SW, FI, RU, NL
  - 2004: EN, FR, RU, PT
  - 2005-06: FR, PT, HU, BG
  - 2007: HU, BG, CZ
  - 2008-09: Persian
  - Both monolingual, bilingual and multilingual evaluation
  - see [Braschler & Peters 2004]

<table>
<thead>
<tr>
<th>Size MB</th>
<th>FR</th>
<th>PT</th>
<th>BG</th>
<th>HU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Docs</td>
<td>487 MB</td>
<td>564 MB</td>
<td>213 MB</td>
<td>105 MB</td>
</tr>
<tr>
<td># token/doc</td>
<td>177,452</td>
<td>210,734</td>
<td>69,195</td>
<td>49,530</td>
</tr>
<tr>
<td># queries</td>
<td>50</td>
<td>50</td>
<td>49</td>
<td>50</td>
</tr>
<tr>
<td># rel. doc./query</td>
<td>50.74</td>
<td>58.08</td>
<td>15.88</td>
<td>18.78</td>
</tr>
</tbody>
</table>

Evaluation Campaigns

- **NTCIR** (research.nii.ac.jp/ntcir/)
  - Started in 1999: EN, JA
  - NTCIR-2 (2001): EN, JA, ZH (traditional)

Topic descriptions available in different languages (CLEF 2005)

- EN: Nestlé Brands
  - FR: Les Produits Nestlé
  - PT: Marcas da Nestlé
  - HU: Nestlé márkák
  - BG: Продуктите на Нестле

- EN: Italian paintings
  - FR: Les Peintures Italiennes
  - PT: Pinturas italianas
  - HU: Olasz (itáliai) festmények
  - BG: Итальянски картини
Evaluation Campaigns

- FIRE (www.isical.ac.in/~fire/)
  - Started in 2008
  - 2008, 2009, 2010 Hindi, Bengali and Marathi
  - 2011 Tamil & Gujarati added
  - IR and CLIR, newspapers collections
  - Few resources, noisy data
  - Other languages in the next years (Punjabi, Telugu)

Evaluation Methodology

- Compare retrieval performance using a test collection
- To compare relatively the performance of two techniques:
  - each technique used to evaluate test queries
  - results (set or ranked list) compared using some performance measure
  - most common measures - precision and recall

Effectiveness measure

- MAP Mean Average Precision
- MRR Mean Reciprocal Rank
- Statistical testing is required

Outline

- MLIA Motivation & Evaluation Campaigns
- Indexing
- Translation
- Matching

Indexing

- Step 1: Select, format, coding
- Step 2: Language identification
- Step 3: Granularity (XML)
- Step 4: Segmentation (tokenization)
- Step 5: Normalization (stemmer)
- Step 6: Enrichment
Indexing Step 1: Preprocessing

- Select sources to be indexed
- Ensure proper handling of the source material by subsequent processing steps
- Unify format and coding
- Do necessary pre-processing
  - Various issues: remove duplicates, headers/footers, etc.

What does that means for non-English IR?

Beyond Just English

- Alphabets
  - Latin alphabet (26)
  - Cyrillic (33): спутник
  - Arabic (28), Hebrew
  - Other Asian languages: Hindi, Thai
- Syllabaries
  - Japan: Hiragana (46) における Katakana (46) プラッツ
  - Korean: Hangul (8,200) 정보검색시스템
- Ideograms
  - China (13,000/7,700) 中国人, Japan (8,800) ぼんぼり
- Transliteration/romanization is (sometimes) possible see LOC at www.loc.gov/catdir/cps0/roman.html

Encoding systems

- ASCII is limited to 7 bits
- Windows, Macintosh, BIG5, GB, EUC-JP, EUC-KR, ...
- ISO-Latin-1 (ISO 8859-1 West European), Latin-2 (East European), Latin-3 (South European), Latin-4 (North European), Cyrillic (ISO-8859-5), Arabic (ISO-8859-6), ...
- Unicode (UTF-8, see www.unicode.org)
- One language ≠ one encoding
- Input / output devices (at least the query)
- Tools
  - What is the result of a sort on Japanese words?
Even English is not Just English

- Historical variations in English
  Our Father, who is in heaven, may your name be kept holy. May your kingdom come into being. May your will be followed on earth, as it is in heaven.
- Around 1600
  Our Father which are in heaven, hallowed be thy Name. Thy kingdom come. Thy will be done, on earth as it is in heaven.
- Around 1400
  Oure fadir that art in heuenes halowid be thi name, thy kyngdom come to, be thi will don in erthe es in heuene,
- Around 1000
  Faeder ure the eart on heofonum, si thin nama gehalgod. Tobecume thine rice. Gewurthe in willa on eorstan ssa swa on heofonum.

Indexing Step 2: Identification

- Most of the following steps are language dependent
- It is necessary to identify the language of the text to be processed
  - on document level
  - on paragraph level, or
  - on sentence level
- Language identification (common words, frequencies of bigrams, trigrams, …)

Language Identification

- Is important (see EuroGov at CLEF 2005)
  - Important to apply the appropriate stopword / stemmer
  - the same language may used different coding (RU)
  - the same information could be in available in different languages
- Domain name does not always help
  - in .uk, 99.05% are written in EN
  - in .de, 97.7% in DE (1.4% in EN, 0.7% in FR)
  - in .fr, 94.3% in FR (2.5% in DE, 2.3% in EN)
  - in .fi, 81.2% in FI (11.5% in SW, 7.3% in EN)
- And multilingual countries and organizations
  - in .be, 36.8% in FR, 24.3% in NL, 21.6% in DE, 16.7 in EN
  - In .eu, ?

Indexing Step 3: Granularity

- What is the granularity of retrieved items?
  - Entire document
  - Sub-document (chapter, paragraph, passage, sentence)
  - Extract only some logical elements (title & abstract)
  - Super-document (aggregation of documents, linked documents, folders)

→ Will not be discussed further (see, e.g., XML IR)
Indexing Step 4: Segmentation

- The document is split into "valid" tokens
  - "To be or not to be" 6 tokens, but 4 word types
- The tokens are suitable to form the index structure
- "Undesirable" tokens are eliminated
  - non-content bearing tokens
  - special characters
  - numbers, date, amounts in $
  - very short or very long tokens, ...

Segmentation

- What is a word / token? Sequence of letters?
  - I'll send you Luca's book
  - C|net & Micro$oft
  - IBM360, IBM-360, ibm 360, ...
  - Richard Brown
  - brown paint
  - Brown is the ...
  - flowerpot
  - flower-pot
  - flo-wer-pot (hyphen ?)

Segmentation

- Compound construction
  - Morphological characteristic used by many languages
    - EN: handgun, viewfinder
    - FR: "porte-clefs" (key ring) "chemin de fer" (railway)
    - IT: "capoufficio" (chief of the office) = "capo" + "ufficio"
      but "capiufficio" (plural)
      but "capogiro" (sing) and "capogiri" (plural) (dizinesss)
    - BU: "радиоапарат" = "radio" (radio) + "апарат" (receiver)
    - Fi: "työviikko" = "työ" (work) + "viikko" (week)
    - HU: "hétvégé" = "hét" (week / seven) + "vég" (end)
  - Compound may have an impact on retrieval effectiveness
Segmentation

- Important in ZH

我不是中国人
我 不 是 中国人
I not be Chinese

- Different segmentation strategies possible
  (longest matching principle, mutual information, dynamic programming approach, morphological analyzer, see MandarinTools (www.mandarintools.com))

Segmentation

- Language independent approach
  n-gram indexing [McNamee & Mayfield 2004], [McNamee 2008]
  different forms possible
  “The White House”
  or → “the”, “whit”, “hite”, “hous”, “ouse”
- usually presents an effective approach when facing with new and less known language
- a classical indexing strategy for JA, ZH or KR
- trunc-n, consider only the first n letters compute → “compu”

Segmentation

A Chinese sentence, various representations

我不是中国人
我 不 是 中国人
I not be Chinese

Unigrams
我不 不 是 中国人

Bigrams
我不 不 是 中 中国 国人

Unigrams and bigrams
我不 不 是 中 中国 国人

Words (MTSeg)
我不是中国人

Segmentation in ZH

ZH: Unigram & bigram > word (MTool) ≈ bigram
n-gram approach (language independent) better than language-dependent (automatic segmentation by MTool) [Abdou & Savoy 2006]
Baseline in bold, difference statistically significant underlined
JA: Unigram & bigram ≈ word (Chasen) ≥ bigram [Savoy 2005]

<table>
<thead>
<tr>
<th>MAP / ZH (T)</th>
<th>NTICIR-5</th>
<th>unigram</th>
<th>bigram</th>
<th>word (MTool)</th>
<th>uni+ bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB2</td>
<td>0.2774</td>
<td>0.3042</td>
<td>0.3246</td>
<td>0.3433</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>0.2955</td>
<td>0.2594</td>
<td>0.2800</td>
<td>0.2943</td>
<td></td>
</tr>
<tr>
<td>Okapi</td>
<td>0.2879</td>
<td>0.2995</td>
<td>0.3231</td>
<td>0.3321</td>
<td></td>
</tr>
<tr>
<td>tf idf</td>
<td>0.1162</td>
<td>0.2130</td>
<td>0.1645</td>
<td>0.2201</td>
<td></td>
</tr>
</tbody>
</table>
Stopword List

- Remove non-content bearing tokens
- Frequent and insignificant terms (det., prep., conj., pron.)
- Could be problematic (in French, “or” could be translated by “gold” or “now / thus”), “who” and WHO (World Health Org.) with diacritics too (e.g., “été” = summer / been, but “ete” does not exist).
- May be system-dependent (e.g., a QA system need the interrogative pronoun in the query)
- Could be “query-dependent” (remove only words that appear frequently in the topic formulation) (see TLR at NTCIR-4)

Stopword List

- For the English language
- No clear and precise decision rule
- Intelligent matching between query & document terms
- Reduce the size of the inverted file (30% to 50%)
- The SMART system suggests 571 words (e.g., “a”, “all”, “are”, “back”, “your”, “yourself”, “years”…)
- Fox [1990] suggests 488 terms
- The DIALOG system suggests 9 terms (“an”, “and”, “by”, “for”, “from”, “of”, “the”, “to”, “with”) due to problem with query “vitamin a” or “IT engineer”
- WIN system (Thomson Reuters) uses one term (“the”)

Stopword List


<table>
<thead>
<tr>
<th>MAP</th>
<th>SMART (571 words)</th>
<th>Short (9 words)</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Okapi</td>
<td>0.4516</td>
<td>0.4402</td>
<td>0.3839</td>
</tr>
<tr>
<td>DFR-I(n)B2</td>
<td><strong>0.4702</strong></td>
<td><strong>0.4743</strong></td>
<td><strong>0.4737</strong></td>
</tr>
<tr>
<td>DFR-PL2</td>
<td>0.4468</td>
<td>0.4463</td>
<td>0.3159</td>
</tr>
<tr>
<td>DFR-PB2</td>
<td>0.4390</td>
<td>0.3258</td>
<td>0.0267</td>
</tr>
<tr>
<td>tf idf</td>
<td>0.2742</td>
<td>0.2535</td>
<td>0.2293</td>
</tr>
</tbody>
</table>

Underlined: significant difference with SMART

Stopword List

- Topic #136 (“Leaning Tower of Pisa”, 1 relevant item)
  - AP = 1.0 with SMART stopword list
  - AP = 0.0 with “None” (no stopword list)
  - Presence of many stopwords (e.g., “of,” “the,” “is,” “what”) ranked many non-relevant documents higher than the single relevant.
- Topic #104 (“Super G Gold medal”)
  - AP = 0.4525 when using the SMART stopword list
  - AP = 0.6550 with “None” (no stopword list)
  - The search term “G” included in the stopword list was removed during the query processing.
Indexing Step 5: Normalization

- Tokens are normalized in order to reach features which are suitable for retrieval
- This is one objective of the use of a controlled vocabulary in manual indexing
  - normalize orthographic variations (e.g., "judgment" or "judgement")
  - Case normalization (e.g., Moon vs. moon)
  - lexical variants (e.g., "analyzing", "analysis")
  - equivalent terms that are synonymous in meaning (e.g., "film", "movie")

Normalization

- Diacritics differ from one language to another ("résumé", “Äpfel”)
- could be used to distinguish the meaning (e.g., "tache" (task) or "tâche" (mark, spot))
- Normalization / Proper nouns
  - Spelling may change with languages
    - Gorbachev, Gorbachew, Gorbachov
    - Mona Lisa ↔ La Joconde ↔ La Gioconda
- Specialized thesauri are useful
  - Unified List of Artist Names
  - Thesaurus of Geographic Names
- Think about SMS language (BTW, 4Y, P2P, …)

Normalization

- Stemming
- Inflectional (light)
  - number (sing / plural) horse, horses
  - gender (femi / masc …) actress, actor
  - grammatical case Paul’s verbal forms (person, tense), jumping, jumped relatively simple in English (‘-s’, ‘-ing’, ‘-ed’)
- derivational (stem + suffix = word)
  - forming new words (changing POS)
  - ‘-ably’, ‘-ment’, ‘-ship’
  - admit → {admission, admittance, admittedly}
Stemming

- Algorithmic Stemmer (rule-based)
  - Lovins (1968) → 260 rules
  - Porter (1980) → 60 rules
  - S-stemmer [Harman 1991] → 3 rules
- concentrate on the suffixes
- add quantitative constraints
- add qualitative constraints
- rewriting rules
- IR is usually based on an average IR performance
- Over-stemming or under-stemming are possible

"organization" → "organ"

Evaluation CLEF 2001 to CLEF 2006 (LA Times (94) & Glasgow Herald (95)), for 169,477 documents, 284 TD queries

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>S-stem</th>
<th>Porter</th>
<th>Lovins</th>
<th>SMART</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Okapi</td>
<td>0.4345</td>
<td>0.4648</td>
<td>0.4706</td>
<td>0.4560</td>
<td>0.4759</td>
<td>0.4633</td>
</tr>
<tr>
<td>PL1</td>
<td>0.4251</td>
<td>0.4553</td>
<td>0.4604</td>
<td>0.4499</td>
<td>0.4624</td>
<td>0.4608</td>
</tr>
<tr>
<td>In,IC2</td>
<td>0.4329</td>
<td>0.4658</td>
<td>0.4721</td>
<td>0.4565</td>
<td>0.4783</td>
<td>0.4671</td>
</tr>
<tr>
<td>LM</td>
<td>0.4240</td>
<td>0.4493</td>
<td>0.4555</td>
<td>0.4389</td>
<td>0.4568</td>
<td>0.4444</td>
</tr>
<tr>
<td>If/Ilf</td>
<td>0.2669</td>
<td>0.2811</td>
<td>0.2839</td>
<td>0.2650</td>
<td>0.2860</td>
<td>0.2778</td>
</tr>
<tr>
<td>Average</td>
<td>0.4291</td>
<td>0.4588</td>
<td>0.4647</td>
<td>0.4503</td>
<td>0.4685</td>
<td>0.4597</td>
</tr>
<tr>
<td>%change</td>
<td>+6.9%</td>
<td>+8.3%</td>
<td>+4.9%</td>
<td>+9.2%</td>
<td>+7.1%</td>
<td></td>
</tr>
</tbody>
</table>

underlined: significant with the best (column)
† with "None"
‡ with "SMART" [Fautsch & Savoy, 2009]

Example

- IF ("*ing") → remove –ing
e.g., "king" → "k", "running" → "runn"
- IF ("*ize") → remove –ize
e.g., "seize" → "se"

To correct these rules:
- IF ("*ing") & (length>3) → remove –ing
- IF ("*ize") & (!final(-e)) → remove –ize
- IF (suffix & control) → replace ...
  "runn" → "run"

Stemming

- Topic #306 ("ETA Activities in France", 1 relevant item)
  - AP = 0.333 without stemming
  - AP = 1.0 with the S-stemmer
  - The term "activities" which after stemming is reduced to "activity". The relevant document contains "activity" three times and "activities" two times.
- Topic #180 ("Bankruptcy of Barings")
  - AP = 0.7652, without stemming
  - AP = 0.0082 when using the SMART stemmer
  - The word "Barings" was stemmed to "bare" (hurt the retrieval performance).
Stemming

Light stemming for other languages?
Usually “simple” for Romance language family
- Example with Portuguese / Brazilian
  Plural forms for nouns → -s (“amigo”, “amigos”)
  but other possible rules (“mar”, “mares”, …)
  Feminine forms -o → -a (“americano” → “americana”)
- Example with Italian
  Plural forms for nouns
  -e → -e (“cani”, “cani”)
  -a → -e (“rosa”, “rose”), …
  Feminine forms -o → -a (“amico” → “amica”)

Stemming (Czech)

- Seven grammatical cases, even for names

<table>
<thead>
<tr>
<th>Case</th>
<th>Paris</th>
<th>Praha</th>
<th>France</th>
<th>Ann</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominative</td>
<td>Paříž</td>
<td>Praha</td>
<td>Francie</td>
<td>Anna</td>
</tr>
<tr>
<td>genitive</td>
<td>Paříž</td>
<td>Praž</td>
<td>Francie</td>
<td>Anně</td>
</tr>
<tr>
<td>dative</td>
<td>Paříž</td>
<td>Praž</td>
<td>Francie</td>
<td>Anně</td>
</tr>
<tr>
<td>accusative</td>
<td>Paříž</td>
<td>Praž</td>
<td>Franci</td>
<td>Annu</td>
</tr>
<tr>
<td>vocative</td>
<td>Paříž</td>
<td>Praha</td>
<td>Francie</td>
<td>Annu</td>
</tr>
<tr>
<td>locative</td>
<td>Paříž</td>
<td>Prahou</td>
<td>Franci</td>
<td>Anně</td>
</tr>
<tr>
<td>instrumental</td>
<td>Paříž</td>
<td>Prahou</td>
<td>Franci</td>
<td>Annou</td>
</tr>
</tbody>
</table>

Stemming

More complex for Germanic languages
- Various forms indicate the plural (+ add diacritics)
  “Motor”, “Motoren”; “Jahr”, “Jahre”;
  “Apfel”, “Äpfel”; “Haus”, “Häuser”
- Grammatical cases imply various suffixes
  (e.g., genitive with “-es” “Staats”, “Mannes”)
  and also after the adjectives (“einen guten Mann”)
- 3 genders x 2 numbers x 4 cases = 24 possibilities!
- Compound construction
  (“Lebensversicherungs-gesellschaftsangestellter”
   = life + insurance + company + employee)

Stemming

Mean relative improvement due to (light) stemming
+4% with the English language
+4% Dutch
+7% Spanish
+9% French
+15% Italian
+19% German
+29% Swedish
+34% Bulgarian
+40% Finnish
+44% Czech
Decompounding (German)

Given a set of words (no stemming, but upper → lower) with their frequencies in a corpus:

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer</td>
<td>2452</td>
</tr>
<tr>
<td>computers</td>
<td>79</td>
</tr>
<tr>
<td>sicherheit</td>
<td>6583</td>
</tr>
<tr>
<td>sicher</td>
<td>4522</td>
</tr>
<tr>
<td>bank</td>
<td>9657</td>
</tr>
<tr>
<td>bund</td>
<td>7032</td>
</tr>
<tr>
<td>bundes</td>
<td>2884</td>
</tr>
<tr>
<td>bundesbank</td>
<td>1453</td>
</tr>
<tr>
<td>präsident</td>
<td>24041</td>
</tr>
<tr>
<td>port</td>
<td>1091</td>
</tr>
<tr>
<td>ports</td>
<td>2</td>
</tr>
<tr>
<td>sport</td>
<td>1483</td>
</tr>
<tr>
<td>winter</td>
<td>1643</td>
</tr>
<tr>
<td>winters</td>
<td>148</td>
</tr>
<tr>
<td>wintersport</td>
<td>44</td>
</tr>
<tr>
<td>wintersports</td>
<td>2</td>
</tr>
</tbody>
</table>

Decompounding (German)

Try with “Bundesbankpräsident”

“bundesbank” 1453 / “präsident” 24041

“bund” 7032 / ’es’ /
“bank” 9657

A similar issue with compounds also exists in other Germanic languages, such as Dutch, Swedish, ... as well as other languages (Hungarian)

Indexing Step 6: Enrichment

Documents are enriched with extra features, or with more specialised features

- Named Entity recognition
- Thesauri for expansion
- Anchor text from inlinks
- Contextual information (from user profiles, from linked pages, from clustering, ...)
- ...

Outline

- MLIR Motivation & Evaluation Campaigns
- Indexing
- Translation
- Matching
Translation

Difficult problem, even for humans

- Cairo, Egypt
  "Unaccompanied ladies not admitted unless with husband or similar"
- On a Japanese medicine bottle,
  "Adults: 1 tablet 3 times a day until passing away"


Manual translation is the norm
- 1,200 persons are working for the Translation Bureau in Ottawa
- Directorate-General for Translation (DGT) (EU) with around 2,500 persons (€800 M)

Translation Problem

- Not a word-by-word translation, but translate the meaning
- "horse" = "cheval"?
  - yes (a four-legged animal)
    "horse-race" = course de chevaux
  - yes in meaning, not in the form
    "horse-show" = "concours hippique"
    "horse-drawn" = "hippomobile"
- different meaning / translation
  "horse-fly" = "taon"
  "horse sense" = "gros bon sens"
  "to eat like a horse" = "manger comme un loup"

Translation Ambiguity

- "post"
  Mail? Post office
  Position? Academic post
  Pole? A long and straight stick
  Other? An entry in a blog, pillar,
  a structural element of a car, a military base,
  a passing route in American football,
  post-mortem examination,
  Post Emily (1873-1960),
  Washington Post,
  Post Records (US label)

Automatic Translation

- In general: IR performance from 50 to 75% of the equivalent monolingual case (TREC-6)
  up to 80% to 100% (CLEF 2005)
- Do we need to present (to the user) the translation?
  - yes: to summarize a result
  - no: simple bag-of-words (sent to the IR process)
- Can the user help (translating / selecting)?
  - "I'm not an expert but I can recognize the correct translation of a painting / artist name in Italian"
Automatic Translation
- In many cases, the context could be rather short
- Query translation could be a mix of bag-of-words and phrase
  E.g., “orange plate with a table”
  difficult to understand/classify
  “orange plate” a noun phrase or a bag of words
- Legend of statistical tables
- Caption of images
- Short description of a cultural object
  (with a mixed of languages, e.g., The European Library)

Translation Strategies
- Ignore the translation problem!
  Sentence in one language is misspelled expression of the other (near cognates) and with some simple matching rules, a full translation is not required
  (e.g., Cornell at TREC-6, Berkeley at NTCIR-5)
- Machine-readable bilingual dictionaries (MRD)
  provide usually more than one translation alternatives
  (take all? the first?, the first k? same weight for all?)
- OOV problem (e.g., proper noun)
  could be limited to simple word lists
  Must provide the lemmas (not the surface words!)
  (relatively easy with the English language)

OOV
- Out-Of-Vocabulary
- Dictionary has a limited coverage (both in direct dictionary-look up or within an MT system)
- Occurs mainly with names (geographic, person, products)
- The correct translation may have more than one correct expression (e.g. in ZH)
- Using the Web to detect translation pairs, using punctuation marks, short context and location (e.g. in EN to ZH IR) [Y. Zhang et al. TALIP]
- Other approaches to improve the translation?

Translation Strategies
- Machine translation (MT)
  - various off-the-shelf MT systems available
  - quality (& interface) varies across the time
- Statistical translation models [Nie et al. 1999]
  - various statistical approaches suggested
- MOSES statistical machine translation model
  www.statmt.org/moses/
- Statistical translation methods tend to dominate the field
- How can we improve the translation process?
Pre-Translation Expansion

- Idea: Add terms into the query before translating it. [Ballesteros & Croft, 1997]
  - The submitted request is usually short.
  - Ambiguity could be high
  - Usually improve the retrieval effectiveness (e.g., Rocchio)
- Good example:
  - Topic #339 "Sinn Fein and the Anglo-Irish Declaration."
    - political british street party anglo-irish declaration britain adam sinn irish ireland government leader fein anglo talk peace northern downing ira"
- Useful additional terms could be morphological related terms (British, Britain, UK)

Pre-Translation Expansion

- More problematic example:
  - Topic #268 "Human Cloning and Ethics."
    - Expanded query
      "parent called call victim human mobile phone made year development fraud ethic cloned time number research stolen cloning clone embryo"
- The problem?
  - We add related terms not semantically related but statistically (according to the target collection)
  - Similar corpus, similar period (e.g., names), similar countries, similar thematic;

Cultural Difference

- The same concept may have different translation depending on the region / country / epoch
  - E.g. "Mobile phone"
    - « Natel » in Switzerland
    - « Cellulaire » in Quebec
    - « Téléphone portable » in France
    - « Téléphone mobile » in Belgium

Automatic Translation (Example)

- "Death of Kim Il Sung"
  - Manually "Mort de Kim Il Sung"
  - Systran "La mort de Kim Il chantée"
  - Babylon "mort de Kim Il chanter"
  - Babylon "Tod von Kim ilinium singen"
- "Who won the Tour de France in 1995?"
  - Manually "Qui a gagné le tour de France en 1995"
  - Systran "Organisation Mondiale de la Santé, le, France"
Automatic Translation (Example)

- Example EN → IT (idiomatic)

Translation

A better translation does not always produce a better IR performance!

<table>
<thead>
<tr>
<th>Translation</th>
<th>Query</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverso</td>
<td>Invasion der Vereinigten Nationen Vereinigter Staaten Haitis. Finden Sie Dokumente auf der Invasion Haitis durch Vereinte Nationen Vereinigte Staaten Soldaten.</td>
<td>40.07</td>
</tr>
<tr>
<td>Free</td>
<td>U.N. UNS Invasion von Haiti. Fund dokumentiert auf der Invasion von Haiti durch U.N. UNS Soldaten</td>
<td>72.14</td>
</tr>
</tbody>
</table>

Translation

On a large query set (284 CLEF 2001-06, English corpus)
Original query written in English (Title-only) [Dolamic & Savoy 2010b]
Statistical significant difference (*)

<table>
<thead>
<tr>
<th>MAP</th>
<th>Mono</th>
<th>From ZH</th>
<th>From DE</th>
<th>From FR</th>
<th>From SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>l(ne)C2</td>
<td>0.4053</td>
<td>0.3340*</td>
<td>0.3618*</td>
<td>0.3719*</td>
<td>0.3741*</td>
</tr>
<tr>
<td>Okapi</td>
<td>0.4044</td>
<td>0.3327*</td>
<td>0.3625*</td>
<td>0.3692*</td>
<td>0.3752*</td>
</tr>
<tr>
<td>LM</td>
<td>0.3708</td>
<td>0.3019*</td>
<td>0.3305*</td>
<td>0.3400*</td>
<td>0.3426*</td>
</tr>
<tr>
<td>tf idf</td>
<td>0.2392*</td>
<td>0.1920*</td>
<td>0.2266*</td>
<td>0.2294*</td>
<td>0.2256*</td>
</tr>
<tr>
<td>diff</td>
<td>-18.2%</td>
<td>-9.3%</td>
<td>-7.3%</td>
<td>-7.1%</td>
<td></td>
</tr>
</tbody>
</table>

Translation

Original query written in English (284 T-only)
Automatic translation done by Google (May 2007)
Statistical significant difference (*) [Dolamic & Savoy 2010b]

<table>
<thead>
<tr>
<th>MAP</th>
<th>Mono</th>
<th>From ZH</th>
<th>From DE</th>
<th>From FR</th>
<th>From SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>l(ne)C2</td>
<td>0.4053</td>
<td>0.3340*</td>
<td>0.3618*</td>
<td>0.3719*</td>
<td>0.3741*</td>
</tr>
<tr>
<td>Okapi</td>
<td>0.4044</td>
<td>0.3327*</td>
<td>0.3625*</td>
<td>0.3692*</td>
<td>0.3752*</td>
</tr>
<tr>
<td>LM</td>
<td>0.3708</td>
<td>0.3019*</td>
<td>0.3305*</td>
<td>0.3400*</td>
<td>0.3426*</td>
</tr>
<tr>
<td>tf idf</td>
<td>0.2392*</td>
<td>0.1920*</td>
<td>0.2266*</td>
<td>0.2294*</td>
<td>0.2256*</td>
</tr>
<tr>
<td>diff</td>
<td>-18.2%</td>
<td>-9.3%</td>
<td>-7.3%</td>
<td>-7.1%</td>
<td></td>
</tr>
</tbody>
</table>
Translation

Original query written in English (284 T-only)
Automatic translation done by Yahoo (May 2007)
Statistical significant difference (*) [Dolamic & Savoy 2010b]

<table>
<thead>
<tr>
<th>MAP</th>
<th>Mono</th>
<th>From ZH</th>
<th>From DE</th>
<th>From FR</th>
<th>From SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>I(ne)C2</td>
<td>0.4053</td>
<td>0.2286*</td>
<td>0.2951*</td>
<td>0.3322*</td>
<td>0.2897*</td>
</tr>
<tr>
<td>Okapi</td>
<td>0.4044</td>
<td>0.2245*</td>
<td>0.2917*</td>
<td>0.3268*</td>
<td>0.2867*</td>
</tr>
<tr>
<td>LM</td>
<td>0.3708</td>
<td>0.2000*</td>
<td>0.2636*</td>
<td>0.3006*</td>
<td>0.2600*</td>
</tr>
<tr>
<td>tf idf</td>
<td>0.2392</td>
<td>0.1289*</td>
<td>0.1846*</td>
<td>0.2065*</td>
<td>0.1812*</td>
</tr>
<tr>
<td>diff</td>
<td>-45.1%</td>
<td>-26.7%</td>
<td>-17.5%</td>
<td>-27.9%</td>
<td></td>
</tr>
</tbody>
</table>

Translation Strategies

Some findings
- The quality (IR view) of MT system has a large variability
- Some languages are more difficult than other (ZH)
- The easiest language is not always the same SP for Google, clearly FR for Yahoo!
- For some IR model and language pair, the difference in MAP could be small
  Google, FR as query language: 0.2392 vs. 0.2294 (-4.1%)

Translation

Where are the real translation problems?
For Google MT system

<table>
<thead>
<tr>
<th>Source</th>
<th>ZH</th>
<th>DE</th>
<th>FR</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>21</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>polysemy</td>
<td>16</td>
<td>4</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>morphology</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>compound</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>other</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Outline

- MLIR Motivation & Evaluation Campaigns
- Indexing
- Translation
- Matching
Matching: Assumptions

- The matching stage needs to assign weights to query (and document) terms
- Remember: we should not require exact matches
- Assumptions:
  - Texts having similar vocabulary tend to have the same meaning
  - More query terms match → more relevant
  - Query terms more frequent in doc → more relevant
  - Rare query terms match → more relevant
  - Query terms clustered tightly in doc → more relevant
  - + others (frequent inlinks, occurrence in title, etc.)

Multilingual IR

- If I need to add one language?
- Bilingual IR, simply translate the query (QT)
- Maybe the "simplest scenario"
- We add query translation to a monolingual IR system
- How to integrate the translation step into the overall system?
  - No translation
  - Only with closely-related languages / writing systems
  - Very limited in multilingual application
    - (proper names, places / geographic names)

One Possible MLIR Flow (QT)

MLIR - Query Translation

More complex matching function can be used.
Including the translation probability \( P(t_q|t_d) \) [Xu et al. 2001], [Kraaij 2004] with Q (and C) written in the source language and D in the target language, we obtain

\[
P(Q \mid D) = \prod_{t_q \in Q} \left( (1 - \alpha) \cdot P(t_q|C) + \alpha \cdot \sum_{t_d \in D} P(t_d|D) \cdot P(t_q|t_d) \right)
\]

How to estimate \( P(t_q|t_d) \) or \( P(s|t) \) the probability of having the term \( s \) in the source language given the term \( t \) in the target language?
(see [Gale & Church 1993], [Nie et al. 1999])
MLIR - Query Translation

\[ p[s|t] = \frac{|\{(S,T)|s \in S \text{ and } t \in T\}|}{|\{T|t \in T\}|} \]

with (S,T) sentence pairs in the corresponding languages, and s, t, the words. We consider all sentence pairs (S,T) having the corresponding terms s and t, and we divide by the number of sentences (in T) containing term t [Kraaij 2004]. Variant Model 1 of IBM [Brown et al. 1993]

Moreover, the corpus C (in the source language) could be different (thematic, time, geographic, etc.) than the corpus in the target language (used by the D and denoted C). We may estimate as:

\[ P[s \mid C] = \sum_{t \in C} P[s \mid t] \cdot P[t \mid C] \]

MLIR – Query Translation

- If I need to consider more than one language?
  - More complex setup
  - A series of bilingual steps
    - Query translation (QT) and search into the different languages, then merging
    - Translate the query into different languages
    - Perform a search separately into each language
    - Merge the result lists

Multilingual IR

Merging problem

<table>
<thead>
<tr>
<th>Document</th>
<th>READING</th>
<th>RATING</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 EN120</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>2 EN200</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>3 EN050</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>4 EN705</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>1 RU050</td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td>2 RU005</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td>3 RU120</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>4 ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multilingual IR
- Round-robin
- Raw-score merging
  \[ \text{Score}_j(D_i) \] document score computed with IR system \( j \)
  \[ RSV(D_i) \] final document score
  \[ RSV(D_i) = \sum_{j=1}^{k} \text{Score}_j(D_i) \]
- Normalize (e.g., by the score of the first retrieved doc = max)
  \[ RSV(D_i) = \sum_{j=1}^{k} \frac{\text{Score}_j(D_i)}{\text{Score}_{\text{Max}}_j} \]

Multilingual IR
- Biased round-robin
  select more than one doc per turn from better ranked lists
- Z-score
  computed the mean and standard deviation
  \[ RSV(D_i) = \sum_{j=1}^{k} \frac{\text{Score}_j(D_i)}{\sigma_j} \]
  with
  \[ \text{Score}_j(D_i) = \frac{(\text{Score}_j(D_i) - \mu_j + \delta_j)}{\sigma_j} \]
- Logistic regression [Le Calvé 2000], [Savoy 2004]
  \[ \text{Score}_j(D_i) = \frac{1}{1 + e^{-(\beta_{j} + \beta_{\text{rank}}(D_i) + 2\lambda RSV(D_i))}} \]

MLIR – Document Translation
- More than two languages
  Why not translating the documents?
- All documents are translated into a single language
- Caveat: what happens if many query languages are possible?
  → combination with query translation, interlingua
- No need for merging step!

<table>
<thead>
<tr>
<th>EN→[EN, FR, FI, RU]</th>
<th>Cond. A</th>
<th>Cond. C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round-robin</td>
<td>0.2386</td>
<td>0.2358</td>
</tr>
<tr>
<td>Raw-score</td>
<td>0.0642</td>
<td>0.3067</td>
</tr>
<tr>
<td>Norm (max)</td>
<td>0.2899</td>
<td>0.2646</td>
</tr>
<tr>
<td>Biased RR</td>
<td>0.2639</td>
<td>0.2613</td>
</tr>
<tr>
<td>Z-score</td>
<td>0.2669</td>
<td>0.2867</td>
</tr>
<tr>
<td>Logistic</td>
<td><strong>0.3090</strong></td>
<td><strong>0.3393</strong></td>
</tr>
</tbody>
</table>
Multilingual IR

- Create a common index using document translation (DT) (see Berkeley CLEF-2003)
- Build an index with all docs translated into a common interlingua (EN for Berkeley at CLEF-2003)
- Search into the (large) index and obtain the single result list
- Mix QT and DT (Berkeley at CLEF 2003, Eurospider at CLEF 2003) [Braschler 2004]
- Variant: Create a multilingual index (see Berkeley TREC-7)
  - Build an index with all docs (written in different languages)
  - Translate the query into all languages
  - Search into the (multilingual) index and thus we obtain directly a multilingual merged list

Conclusion

- Search engines are mostly language independent
- Monolingual
  - stopword list, stemmer, compound construction
  - more morphological analysis could clearly improved the IR performance [FI]
  - tokenization is a problem (ZH, JA)
- Multilingual
  - various translation tools for some pairs of language (EN)
  - more problematic for less-frequently used languages
  - IR performance could be relatively close to corresponding monolingual run
  - merging is not fully resolved (see CMU at CLEF 2005)
Conclusion

- "In theory, practice and theory are the same, but in practice they are not."
  David Hawking, Chief Scientist Funnelback

- The various experiments shown that query-by-query analysis is an important step in scientific investigations. We really need to understand why IR system may (will) fail for some topics. Learn by experiences.

- The real problems (implementation) are crucial
  (Der Teufel liegt im Detail)

  David Crystal

References

- Conference
  - ACM-SIGIR
  - ECIR
  - AIRS

- Journal
  - Information Retrieval Journal, IRJ (Springer)
  - Information Processing & Management, IP&M (Elsevier)
  - Journal of the American Society for Information Science & Technology, JASIST (Wiley)

- Evaluation campaigns: CLEF, NTCIR, TREC, FIRE

References


References


References