Beyond just English Cross-Language IR

J. Savoy
University of Neuchatel
iiun.unine.ch

http://www.clef-campaign.org
http://research.nii.ac.jp/ntcir/
http://trec.nist.gov (TREC-3 to TREC-12)

The 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." [D. Oard & D. Hull, AAAI Symposium on Cross-Language IR, Spring 1997, Stanford]

Outline

- **Motivation and evaluation campaigns**
- Beyond just English, monolingual IR (segmentation & stemming)
- Language identification
- Translation problem
- Translation strategies (bilingual IR)
- Multilingual IR

Motivation

- Facts (www.ethnologue.com)

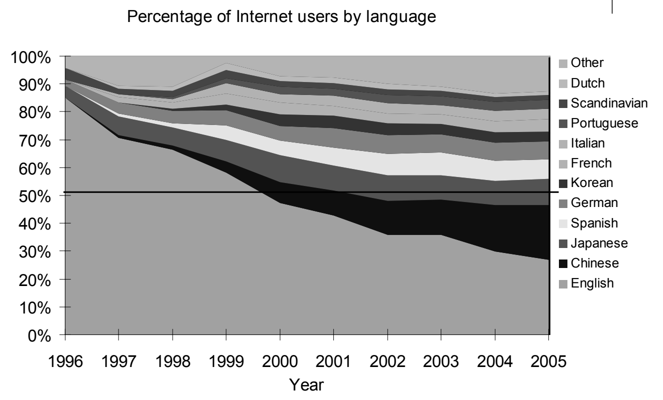
6,800 living languages in the world,
  2,197 in Asia
  2,092 in Africa
  1,310 in Pacific
  1,002 in America
    230 in Europe.
600 of them are writing

80% of the world population speaks 75 different languages
40% of the world population speaks 8 different languages
75 languages are spoken by more than 10 M persons
20 languages are spoken by more than 50 M persons
8 languages are spoken by more than 100 M persons.

Motivation

- One language is
  - a very complex human construction (but so easy to learn when it’s our mother tongue)
  - 100,000 words
  - 10,000 syntactic rules
  - 1,000,000 semantic elements

Motivation



Motivation

- Bilingual / multilingual
  - Many countries are bi- / multilingual (Canada (2), Singapore (2), India (21), EU (20))
    - Official languages in EU: Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Slovak, Slovene, Spanish, Swedish, Irish, (Bulgarian, Romanian).
      Other languages: Catalan, Galcian, Basque, Welsh, Scottish Gaelic, Russian.
    - Working languages in EU: English, German, French;
      In UN: Arabic, Chinese, English, French, Russian, Spanish.
  - Court decisions written in different languages
  - Organizations: FIFA, WTO, UBS, Nestlé, …

Motivation

- Bilingual / multilingual
  - people may express their needs in one language and understand another
  - we may written 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), *image*, *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).

Outline

- Motivation and evaluation campaigns
- **Beyond just English, monolingual IR (segmentation & stemming)**
- Language identification
- Translation problem
- Translation strategies (bilingual IR)
- Multilingual IR

Evaluation campaigns

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

  See [Harman 2005]

Evaluation campaigns

- CLEF (www.clef-campaign.org)
  - 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, RO(?)
  - Both monolingual, bilingual and multilingual evaluation
  - Other tasks: domain-specific, interactive, Spoken document (2002 →), Image-CLEF (2003 →), QA(2003 →), Web(2005 →), GeoCLEF (2005 →) see [Braschler & Peters 2004]

Evaluation campaigns (CLEF 2005)

	FR	PT	BG	HU
Size MB	487 MB	564 MB	213 MB	105 MB
Docs	177,452	210,734	69,195	49,530
# token/ doc	178	213	134	142
# queries	50	50	49	50
# rel. doc./ query	50.74	58.08	15.88	18.78

Evaluation campaigns

- General topic with large and international coverage
  - « Pension Schemes in Europe »
  - « Brain-Drain Impact »
  - « Football Refereeing Disputes »
  - « Golden Bear »
- More national / regional coverage
  - « Falkland Islands »
  - « Swiss referendums »

Evaluation campaigns

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: Италиански картини

Evaluation campaigns

- NTCIR (research.nii.ac.jp/ntcir/)
  - Started in 1999: EN, JA
  - NTCIR-2 (2001): EN, JA, ZH (traditional)
  - NTCIR-3 (2002): NTCIR-4 (2004), and NTCIR-5 (2005): EN, JA, KR, ZH (traditional) and patent (JA), QA (JA), Web (.jp), Summarization
  - NTCIR-6 (2007): JA, KR, ZH (traditional)

Evaluation campaigns (NTCIR-5)

	EN	JA	ZH	KR
Size MB	438 MB	1,100 MB	1,100 MB	312 MB
Docs	259,050	858,400	901,446	220,374
Coding	ASCII	EUC-JP	BIG5	EUC-KR
# queries	49	47	50	50
# rel. doc./ query	62.73	44.94	37.7	36.58

Beyond just English

<TOPIC>

<TITLE>時代華納，美國線上，合併案，後續影響</TITLE>

<DESC> 查詢時代華納與美國線上合併案的後續影響。</DESC>

<NARR>

<BACK>時代華納與美國線上於2000年1月10日宣佈合併，總市值估計為3500億美元，為當時美國最大宗合併案。</BACK>

<REL>評論時代華納與美國線上的合併對於網路與娛樂媒體事業產生的影響為相關。敘述時代華納與美國線上合併案的發展過程為部分相關。內容僅提及合併的金額與股權結構轉換則為不相關。</REL>

</NARR>

<CONC>時代華納，美國線上，李文，Gerald Levin，合併案，合併及採購，媒體業，娛樂事業</CONC>

</TOPIC>

Beyond just English

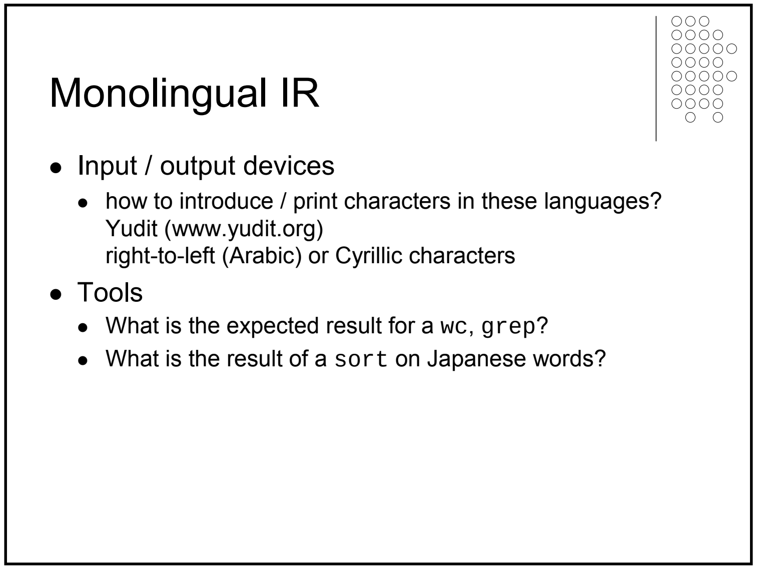
- Other examples
  - Strč prst skrz krk
  - Mitä sinä teet?
  - Mam swoją książkę
  - Nem fáj a fogad?
  - Er du ikke en riktig nordmann?
  - Добре дошли в България!
  - Fortuna caeca est
  - نهارسعيد

Beyond just English

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

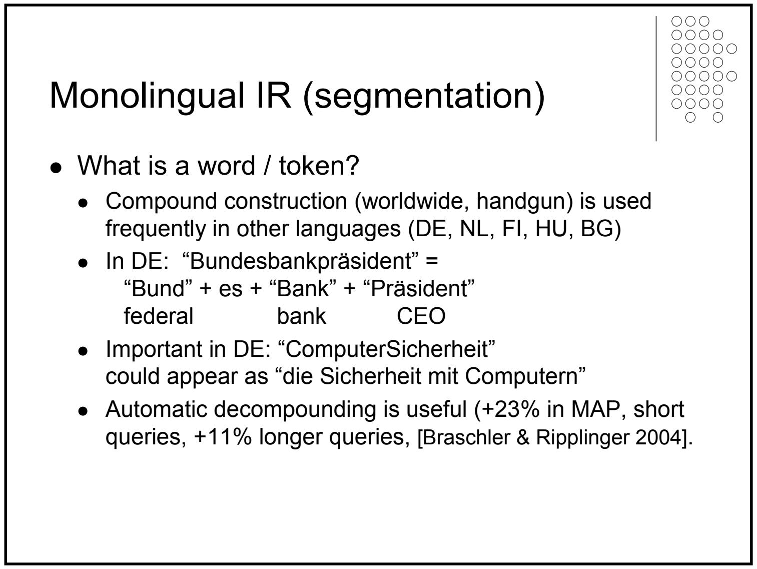
Monolingual IR

- 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), Greek (ISO-8859-7), Hebrew (ISO-8859-8), …
  - Unicode (UTF-8, see www.unicode.org)



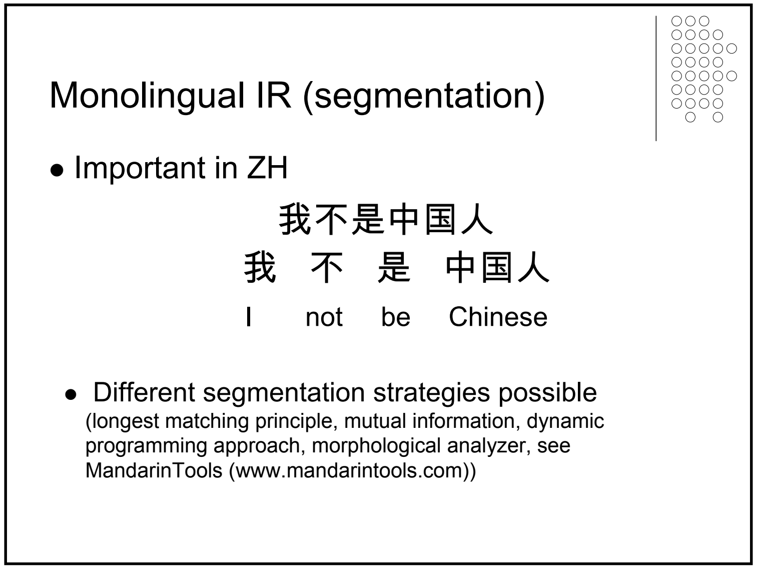
Monolingual IR

- Input / output devices
  - how to introduce / print characters in these languages?
    Yudit (www.yudit.org)
    right-to-left (Arabic) or Cyrillic characters
- Tools
  - What is the expected result for a `wc`, `grep`?
  - What is the result of a `sort` on Japanese words?



Monolingual IR (segmentation)

- What is a word / token?
  - Compound construction (worldwide, handgun) is used frequently in other languages (DE, NL, FI, HU, BG)
  - In DE: “Bundesbankpräsident” =
    “Bund” + es + “Bank” + “Präsident”
    federal bank CEO
  - Important in DE: “ComputerSicherheit”
    could appear as “die Sicherheit mit Computern”
  - Automatic decompounding is useful (+23% in MAP, short queries, +11% longer queries, [Braschler & Ripplinger 2004].



Monolingual IR (segmentation)

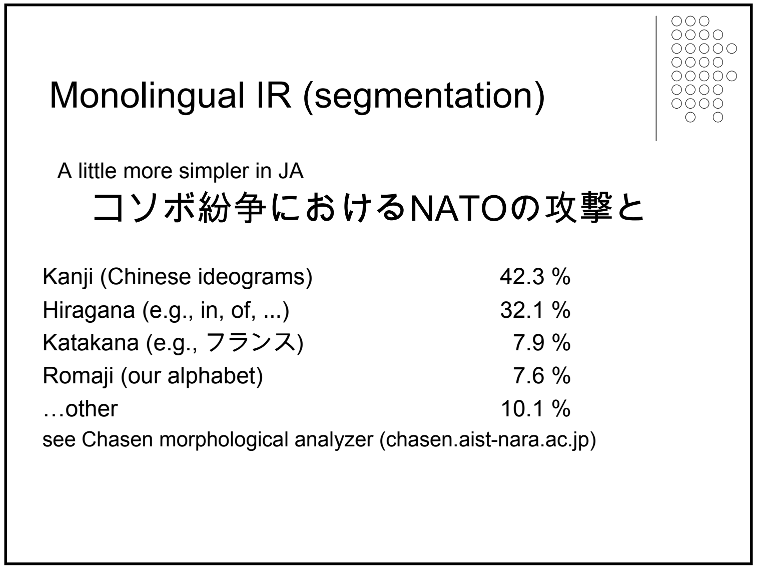
- Important in ZH

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))



Monolingual IR (segmentation)

A little more simpler in JA

コソボ紛争におけるNATOの攻撃と

Kanji (Chinese ideograms)	42.3 %
Hiragana (e.g., in, of, ...)	32.1 %
Katakana (e.g., フランス)	7.9 %
Romaji (our alphabet)	7.6 %
…other	10.1 %

see Chasen morphological analyzer (chasen.aist-nara.ac.jp)

Monolingual IR (segmentation)

The same concept could be expressed by four different compound constructions in KR.

정보 (information) 검색 (retrieval) 시스템 (system)
정보검색 (information retrieval) 시스템 (system)
정보 (information) 검색시스템 (retrieval system)
정보검색시스템

see Hangul Analyser Module (nlp.kookmin.ac.kr)

Monolingual IR

- Language independent approach
  n-gram indexing [McNamee & Mayfield 2004]
  - automatically segment each sentence
  - different forms possible
    “The White House”
    → “The “, “he W”, “h Wh”, “ Whi”, “Whit”, “hite”, …
    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

Monolingual IR

A Chinese sentence

我不是中国人

Unigrams
我 不 是 中 国 人
Bigrams
我不 不是 是中 中国 国人
Unigrams and bigrams
我, 不, 是, 中, 国, 人, 我不, 不是, 是中, 中国, 国人

Words (MTSeg)
我 不 是 中国人

Monolingual IR

A Japanese sentence

クロソフトのWindowsがどのような競合関係

Unigrams
ク ロ ソ フ ト Windows 競 合 関 係
Bigrams
クロ ロソ ソフ フト Windows 競合 合関 関係
Unigrams and bigrams
ク ロ ソ フ ト Windows 競 合 関 係 クロ ロソ ソフ フト
競合 合関 関係
Words (ChaSen)
クロソフト Windows 競合 関係

Monolingual IR

A Korean compound term

정보검색시스템

words

정보검색시스템

Bigrams

정보 보검 검색 색시 시스 스템

Decompounded (HAM)

정보 검색 시스템

Monolingual IR

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

Chinese (T) NTCIR-5	unigram	bigram	word (MTool)	uni+ bigram
PB2	0.2774	**0.3042**	0.3246	0.3433
LM	0.2995	**0.2594**	0.2800	0.2943
Okapi	0.2879	**0.2995**	0.3231	0.3321
tf idf	0.1162	**0.2130**	0.1645	0.2201

Monolingual IR

KR: bigram ≈ HAM > unigram [Abdou & Savoy 2006]

n-gram approach still presents the best performance (not statistically)

Korean (T) NTCIR-5	unigram	bigram	decompound (HAM)
PB2	0.2378	**0.3729**	0.3659
LM	0.2120	**0.3310**	0.3135
Okapi	0.2245	**0.3630**	0.3549
tf idf	0.1568	**0.2506**	0.2324

Monolingual IR

- Diacritics
  - differ from one language to another (“résumé”, “Äpfel”, “leão”)
  - could be used to distinguish the meaning (e.g., “tache” (task) or “tâche (mark, spot))
  - usually related in meaning (e.g., “cure” and “curé” presbytery / parish priest
    however “cure” owns two meanings (as in French)
  - usually there are removed by the IR system (difference in MAP are usually small and non significant)

Monolingual IR

- Normalization / Proper nouns
  - homophones involving proper names. E.g., Stephenson (steam engine), and Stevenson (author) have the same pronunciation in Japanese, Chinese, or Korean languages. Thus both names may be written identically.
  - Spelling may change with languages (Gorbachev, Gorbacheff, Gorbachov)
  - No strict spelling rules (or different spellings possible) E.g., in FR “cow-boy” and “cowboy,” “véto” and “veto,” or “eczéma” and “exéma” (like in English, color, colour, etc.). In DE: different (and contradictory) spelling reforms.

Monolingual IR

- Stopword lists
  - Frequent and insignificant terms (+ pronouns, prep., conj.)
  - Could be problematic (in French, “or” could be translated by “gold” or “now / thus”)
    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 pronouns)
  - Could be “query-dependent” (remove only words that appear frequently in the topic formulation)
    (see TLR at NTCIR-4)

Monolingual IR (stemming)

- Stemming (words & rules)
  - Inflectional
    the number (sing / plural), horse, horses
    the gender (femi / masc), actress, actor
    verbal form (person, tense), jumping, jumped
    relatively simple in English (‘-s’, ‘-ing’, ‘-ed’)
  - derivational
    forming new words (changing POS)
    ‘-ably’, ‘-ment ’, ‘-ship’
    admit → {admission, admittance, admittedly}

Monolingual IR (stemming)

- Stemming
  - with exceptions (in all languages)
    box → boxes, child → children
    one walkman → ? (walkmen / walkmans)
    and other problems: "The data is/are ...", people
  - Suggested approaches (inflection + derivation)
    Lovins (1968) → 260 rules
    Porter (1980) → 60 rules
    Variant: S-stemmer [Harman 1991]: 3 rules
  - Stemming in EN is known [Harman 1991]

Monolingual IR (stemming)

- Based on the grammar
  rule-based (ad hoc approach)
  - concentrate on the suffixes
  - add quantitative constraints
  - add qualitative constraints
  - rewriting rules
- IR is usually based on an average IR performance / could be adapted from specific domain
- Over-stemming or under-stemming are possible
  “organization ” →“organ”

Monolingual IR (stemming)

- 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"

Monolingual IR (stemming)

Light stemming in French (inflectional attached to nouns and adjectives) [Savoy 2004]

Example for the French language (“barons“ → “baron“, “baronnes“ → “baron“)

For words of six or more letters
if final letters are ‘-aux’ then replace ‘-aux’ by ‘-al’,
if final letter is ‘-x’ then remove ‘-x’,
if final letter is ‘-s’ then remove ‘-s’,
if final letter is ‘-r’ then remove ‘-r’,
if final letter is ‘-e’ then remove ‘-e’,
if final letter is ‘-é’ then remove ‘-é’,
if final two letters are the same, remove the final letter

Monolingual IR (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”)

Monolingual IR (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’ “Staates”, “Mannes”)
  and also after the adjectives
  (“einen guten Mann”)
- Compound construction
  (“Lebensversicherungsgesellschaftsangestellter”
  = life + insurance + company + employee)

Monolingual IR (stemming)

Finno-Hungarian family owns numerous cases (18 in HU)

ház	nominative (house)
házat	accusative singular
házakat	accusative plural
házzal	“with” (instrumental)
házon	“over” (superessive)
házamat	my + accusative sing.
házamait	my + accusative + plur.

- In FI, the stem may change (e.g., “matto”, “maton”, “mattoja” (carpet))
  It seems that a deeper morphological analyzer is useful for FI
  (see Hummingbird, CLEF 2004, p. 221-232)
- + Compound construction
  (“internetfüggök”, “rakkauskirje”)

Monolingual IR (stemming)

- Arabic is an important language (TREC-11 / 2002)
- Stemming is important:
  Word = prefix + stem + pattern + suffix
- Stems are three/four letters
  - **ktb** + CiCaC = **k**i**t**a**b**

ki**t**a**b**	a book
ki**t**a**b**i	my book
al**k**i**t**a**b**	the book
ki**t**a**b**uki	your book (femi)
ki**t**a**b**uka	your book (masc)
ka**t**a**b**a	to write
ka**t**i**b**	the writer (masc)
ka**t**i**b**i	the writer (femi)
ma**ktab**	office
ma**ktab**a	library …

  - Spelling variations (for foreign names)
  - The roots are not always the best choice for IR

Monolingual IR (stemming)

Other stemming strategies

- Language usage (vs. grammatical rules)
  or corpus-based stemmer [Xu & Croft 1998]
- Using a dictionary (to reduce the error rate)
  [Krovetz 1993], [Savoy 1993]
- "Ignore" the problem, indexing using *n*-gram
  e.g., "bookshop" → "book" , "ooks", "oksh"
- Effective for ZH, JA, KR …
  [McNamee & Mayfield 2004]

Monolingual IR (stemming)

- Evaluations
- Some experiments in CLEF proceedings
- Other evaluations in [Savoy 2006]
- Main trends (MAP)
  - Stemming > none
  - Differences between stemmer could be stat. significant
  - Simple stemmers for nouns + adjectives tend to perform better, or at the same level of performance than more aggressive stemmers
  - No clear for East Asian languages
    JA: remove Hiragana characters
- Examples in FR

Monolingual IR (stemming)

Stemming is not an error-free procedure

In the query (HU)
"internetfüggők" (internet addiction – person
«függ» is the verb – stem-)

In the relevant documents

"internetfüggőség"	(dependence)	→ "internetfüggőség"
"internetfüggőség**gel**“	(“with“)	→ "internetfüggőség"
"internetfüggőség**ben**“	(“in“)	→ "internetfüggőség"

→ Here the stemming fails

Monolingual IR (stemming)

Based on CLEF-2005 corpus, T queries

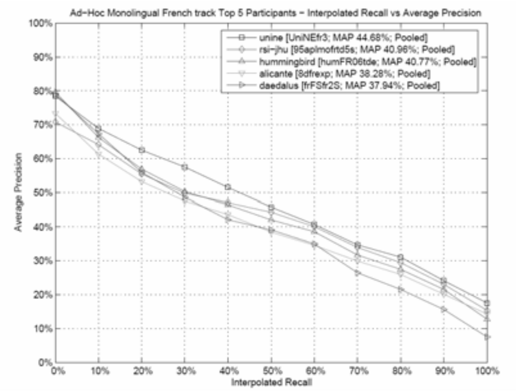
FR (T)	none	UniNE	light ‘-s’	Porter
Okapi	**0.2260**	0.3045	0.2858	0.2978
GL2	**0.2125**	0.2918	0.2739	0.2878
Lnu-ltc	**0.2112**	0.2933	0.2717	0.2808
dtu-dtn	**0.2062**	0.2780	0.2611	0.2758
tf·idf	**0.1462**	0.1918	0.1807	0.1758

Monolingual IR (stemming)

Based on CLEF-2005 corpus, T queries

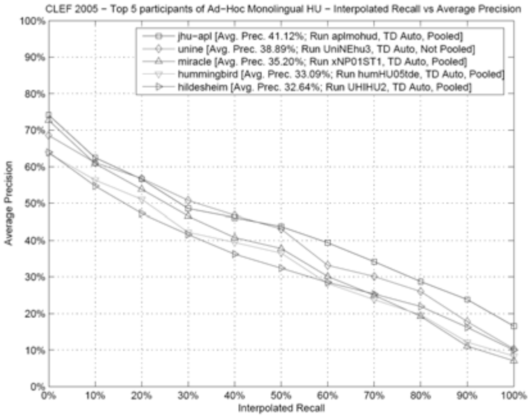
FR (T)	none	UniNE	light ‘-s’	Porter
Okapi	0.2260	0.3045	**0.2858**	0.2978
GL2	0.2125	0.2918	**0.2739**	0.2878
Lnu-ltc	0.2112	0.2933	**0.2717**	0.2808
dtu-dtn	0.2062	0.2780	**0.2611**	0.2758
tf·idf	0.1462	0.1918	**0.1807**	0.1758

Monolingual IR (CLEF 2006)



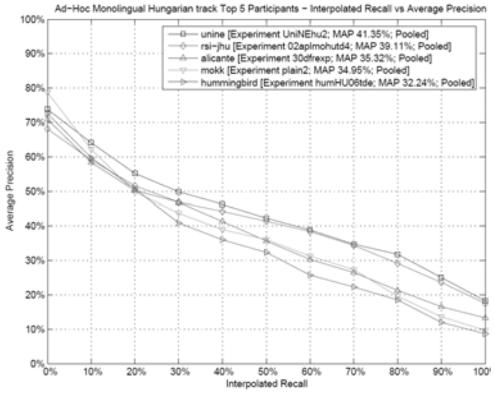
FR, known language
Differences in MAP in the top 5 relatively small
Various IR strategies tend to produce similar MAP

Monolingual IR (CLEF 2005)



- HU, new language
- *n*-gram performs the best
- Improvement is expected (language-dependant)

Monolingual IR (CLEF 2006)



- But it change over time

Outline

- Motivation and evaluation campaigns
- Beyond just English, monolingual IR (segmentation & stemming)
- **Language identification**
- Translation problem
- Translation strategies (bilingual IR)
- Multilingual IR

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`, ?

Language Identification

- Statistics based on
  - short and frequent words
  - trigrams
  - letters distributions
  - gather large number of predictors
- Voting algorithm
  - let each predictor gives its prediction (similarity / distribution distance)
  - maybe: throw away outliers
  - average results

Outline

- Motivation and evaluation campaigns
- Beyond just English, monolingual IR (segmentation & stemming)
- Language identification
- **Translation problem**
- Translation strategies (bilingual IR)
- Multilingual IR

Translation problem

- “non verbum e verbo, sed sensum exprimere de sensu”
- “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 problem

- Loan
  “full-time” → “temps plein”(*)
- Calque
  “igloo” → “iglou”
- Word-by-word translation
  - “a lame duck Congressman” → “canard boiteux”(*)
  - False cognates
    “Requests of Quebec” = “Demandes du Québec”
    “Demands of Quebec” = “Exigences posées par le Québec”
- Translation = equivalence in meaning (not in form “Yield” = “Priorité à gauche” ≠ “Cédez”)

Translation

- “Tainted-Blood Trial”

Manually	“L'affaire du sang contaminé”
Systran	“Épreuve De Corrompu - Sang”
Babylon	“entacher sang procès”

- “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 1995 ”

Outline

- Motivation and evaluation campaigns
- Beyond just English, monolingual IR (segmentation & stemming)
- Language identification
- Translation problem
- **Translation strategies (bilingual IR)**
- Multilingual IR

Automatic translation

- Automatic translation will add ambiguity
  - Multiple translation of each word
  - Use translation probabilities (how?)
  - Query expansion may help
- Require additional and significant language resources
  - Bilingual / multilingual dictionaries (or list of words)
  - Proper names lists
  - Parallel corpora
  - “Compatible corpora” (thematic, time, cultural)
  - MT systems
- Statistical methods dominate the field (SIGIR 2006)

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)
- Topic translation
  - less expensive
- Documents translation
  - done before the search
- Query and documents translation
  - could be very effective
- IR performance from 50 to 75% of the equivalent monolingual case (TREC-6)
  up to 80% to 100% (CLEF 2005)

Translation Strategies

- Machine-readable bilingual dictionaries (MRD)
  - provide usually more than one translation alternatives (take all? the first?, same weight for all?)
  - OOV problem (e.g., proper nouns)
  - could be limited to simple word lists
- 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
  - see project mboi at rali.iro.umontreal.ca/mboi

Translation Strategies

- Pre-translation expansion could be use
  - could be a problem with MT system
- Post-translation expansion
  - usually improve the MAP
- Parallel corpora
  - could be difficult to obtain
  - cultural, thematic and time differences are important
  - the Web could be used
    or more “controlled” source (e.g. Wikipedia)
- “Structured” query could sometimes help [Hedlund *et al.* 2004]
- Better translation of phrases will help
- Evaluation campaigns (specially NTCIR) use a large number of proper names in topic description
  → could be useful to process / translate them with appropriate resource

OOV

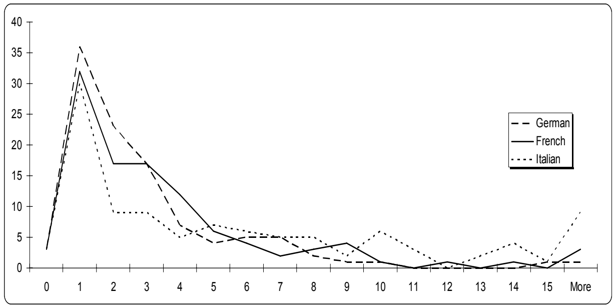
- Out-Of-Vocabulary
  - Dictionary has a limited coverage (both in direct dictionary-lookup or within an MT system)
  - Occurs mainly with names (geographic, person, products)
  - The correct translation may have more than one correct expression (e.g. in ZH)
- Using the Web to detect translation pairs,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

Cultural difference

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

Translation

The number of translation alternatives provided by a bilingual dictionary is usually small (Babylon)



Translation strategies

Example of phrases

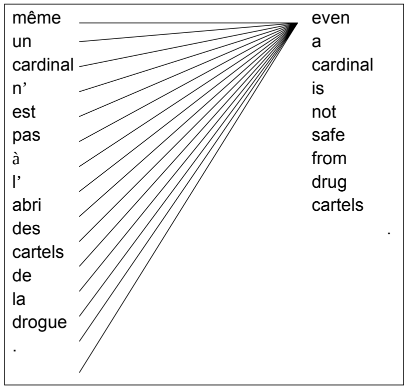
- “Final Four Results”
  - in FR: “final quatre résultat“ (Babylon)
    instead of “Résultats des demi-finales“
  - in DE: “Resultate Der Endrunde Vier “ (Systran)
    instead of “Ergebnisse im Halbfinale“
- “Renewable Power ”
  - in FR, instead of “Energie renouvelable“
    “Puissance Renouvelable“
    “renouvelable pouvoir“
- “Mad Cow Dease ”
  - in FR, instead of “maladie de la vache folle“
    “fou vache malade“
    and the stemming may not find the most appropriate term

Translation strategies

- P[e_j|f_i] is estimated from a parallel training corpus, aligned into parallel sentences [Gale & Church, 1993]
- No syntactic features and position information (IBM model 1, [Brown *et al.*, 1993])
- Process:
  - Input = two sets of parallel texts
  - Sentence alignment *A*: $E_k \leftrightarrow F_l$
  - Initial probability assignment: P[e_j|f_i, *A*]
  - Expectation Maximization (EM): P[e_j|f_i , *A*]
  - Final result: P[e_j|f_i] = P[e_j|f_i , *A*]

Translation strategies

Initial probability assignment P[e_j|f_i, A]

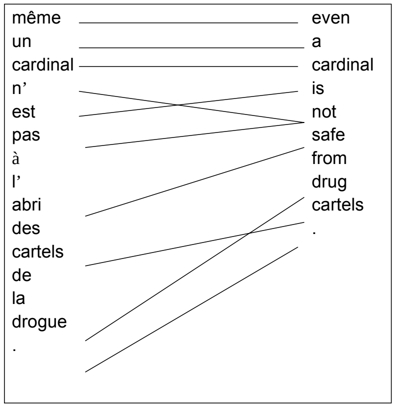


même
un
cardinal
n’
est
pas
à
l’
abri
des
cartels
de
la
drogue
.

even
a
cardinal
is
not
safe
from
drug
cartels
.

Translation strategies

Application of EM: P[e_j|f_i, A]



même
un
cardinal
n’
est
pas
à
l’
abri
des
cartels
de
la
drogue
.

even
a
cardinal
is
not
safe
from
drug
cartels
.

Translation strategies

With parallel corpora [Gale & Church 1991]

- Example with the mboi system (rali.iro.umontreal.ca/mboi)
- “database system”
  - in FR: “(données^0.29472154 base^0.20642714 banque^0.037418656“)
    “système de bases de données“

Translation

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

Translation	Query	MAP
EN (original)	U.N./US Invasion of Haiti. Find documents on the invasion of Haiti by U.N./US soldiers.	
Reverso	Invasion der Vereinter Nationen Vereinigter Staaten Haitis. Finden Sie Dokumente auf der Invasion Haitis durch Vereinte Nationen Vereinigte Staaten Soldaten.	40.07
Free	U N UNS Invasion von Haiti. Fund dokumentiert auf der Invasion von Haiti durch U N UNS Soldaten	72.14

Translation

Comparing 11 different manual translations of the EN queries (T)
[Savoy 2003]

- large variability
- translations provided by CLEF are good (differences are statistically significant, two-tailed, α=5%)

	CLEF	average	max	min
Okapi	0.4162	0.3516	0.4235	0.2929
tf idf	0.2502	0.1893	0.2416	0.0261
binary	0.2285	0.1662	0.2151	0.0288

Translation

Original topics written in EN (Title, Okapi, CLEF-2000)

- automatic translation by Systran
- by Babylon (only the first alternative)
- concatenate both translations

	Manual	Systran	Babylon	Combined
FR word	0.4162	0.2964 (-28.8%)	0.2945 (-29.4%)	0.3314 (-20.4%)
DE 5-gram	0.3164	0.2259 (-28.6%)	0.1739 (-45.1%)	0.2543 (-19.6%)
IT word	0.3398	0.2079 (-38.8%)	0.1993 (-41.3%)	0.2578 (-24.1%)

Translation

Overall statistics may hide irregularities

n same performance that manually translated topic
m automatic translated queries produced better MAP
k manually translated topics achieved better MAP

Language (*n*/*m*/*k*)	Systran	Babylon	Combined
FR (34 queries)	16 / 4 / 14	11 / 3 / 20	11 / 7 / 16
DE (37 queries)	14 / 7 / 16	4 / 5 / 28	6 / 9 / 22
IT (34 queries)	8 / 4 / 22	6 / 4 / 24	0 / 9 / 25

Translation

Could be useful to include the translation process directly into the search formulation.
Starting with a LM [Xu *et al.* 2001]

- Considering a corpus C, a document D and a query Q,
- $P[t_q \mid C]$ probability of the word in the language
- $P[t_q \mid D]$ probability of the word in the document

$$P[Q \mid D] = \prod_{t_q \in Q} [\alpha \cdot P[t_q \mid D] + (1 - \alpha) \cdot P[t_q \mid C]]$$

with $P[t_q \mid D] = \frac{tf\ of\ t_q\ in\ D}{size\ of\ D}$

$$P[t_q \mid C] = \frac{tf\ of\ t_q\ in\ C}{size\ of\ C}$$

Translation

Including the translation probability $P[t_q \mid t_d]$ [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 $P[t_q \mid t_d]$ or $P[s \mid t]$ 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])

Translation

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

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 Model1 of IBM [Brown *et al.* 1993]

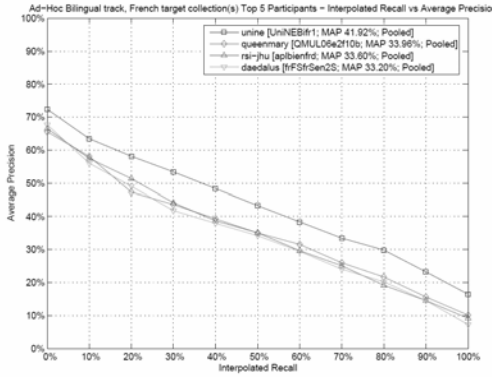
Moreover, the corpus C (in the source language) could be different (thematic, time, geographic, etc.) than the corpus in the target language (used by the D and denoted C$_l$). We may estimate as:

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

Evaluation

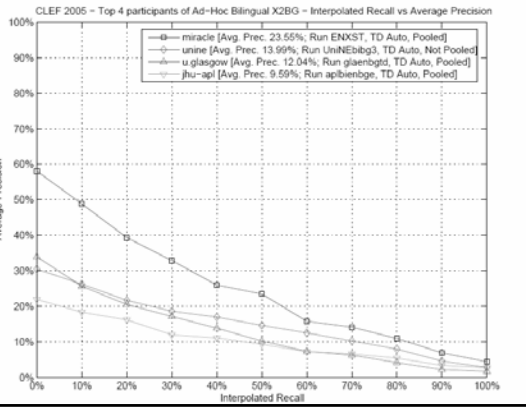
- Different situations are possible
  - Languages may have more or less translation tools / parallel or comparable corpora / morphological tools / IR experiences
  - Languages may be more easier than other
- Direct comparisons between bilingual and monolingual is not always possible
  - Some teams provide runs only for one track
  - Not the same search engines is used for both runs
  - Different settings are used for the monolingual and the bilingual searches

CLIR (CLEF-2006 X → FR)



- Known language
- Various translation tools available
- Track done during five years
- Best mono: 0.4468 (Δ=-6.2%)
- Small difference between the 2nd to the 4th

CLIR (CLEF-2005 X → BG)



New language
Few translation tools available
First year
Best mono: 0.3203 (Δ=-26.5%)
The quality of the translation tool explains the difference between first two runs

Adding new languages

- See CLEF evaluation campaign
  - The *n*-gram approach is language-independent
  - Segmentation & compound construction
  - Diacritics / dialects
  - Coding (unicode?)
  - Stemming (suffixes / prefixes) and some minimal linguistics knowledge
  - Stopword list
- Resource for bilingual IR
  - Bilingual words list
  - Parallel or comparable corpora

Outline

- Motivation and evaluation campaigns
- Beyond just English, monolingual IR (segmentation & stemming)
- Language identification
- Translation problem
- Translation strategies (bilingual IR)
- **Multilingual IR**

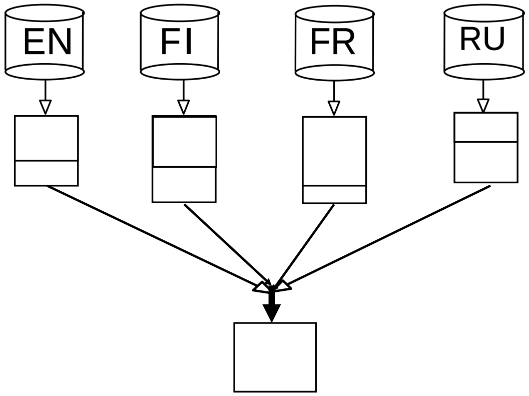
Multilingual IR

- 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
- 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

Multilingual IR

- 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
- Mix QT and DT (Berkely at CLEF 2003, Eurospider at CLEF 2003) [Braschler 2004]
- No translation
  - Only with close languages / writing systems
  - Very limited in multilingual application (proper names, places / geographic names)

Multilingual IR (QT)



Multilingual IR

Merging problem

1	EN120	1.2
2	EN200	1.0
3	EN050	0.7
4	EN705	0.6
...		

1	FR043	0.8
2	FR120	0.75
3	FR055	0.65
4	...	

1	RU050	6.6
2	RU005	6.1
3	RU120	3.9
4	...	

Multilingual IR

- See “Distributed IR”
- Round-robin
- Raw-score merging

$Score_j(D_i)$ document score computed with IR system j

$RSV(D_i)$ final document score

$$RSV(D_i) = \sum_{j=1}^{k} Score_j(D_i)$$

- Normalize (e.g, by the score of the first retrieved doc = max)

$$RSV(D_i) = \sum_{j=1}^{k} Score'_j(D_i)$$
$$with\ Score'_j(D_i) = \frac{Score_j(D_i)}{ScoreMax_j}$$

Multilingual IR

- Biased round-robin

  select more than one doc per turn from better ranked lists)

- Z-score

  computed the mean and standard deviation

$$RSV(D_i) = \sum_{j=1}^{k} Score'_j(D_i)$$
$$with\ Score'_j(D_i) = \frac{(Score_j(D_i) - \mu_j) + \delta_j}{\sigma_j}$$

- Logistic regression [Le Calvé 2000], [Savoy 2004]

$$Score'_j(D_i) = \frac{1}{1 + e^{-[\alpha_j + \beta_{1j} \cdot ln(rank(D_i)) + \beta_{2j} \cdot RSV(D_i)]}}$$

Multilingual IR

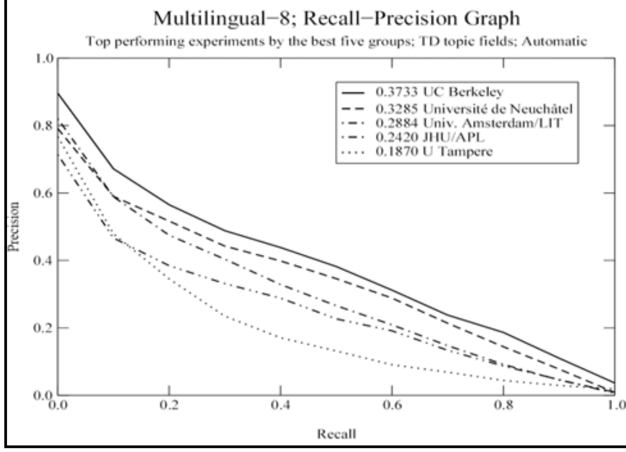
Cond. A best IR system per language (CLEF 2004)
Cond C the same IR system for all languages

EN->{EN, FR, FI, RU}	Cond. A	Cond. C
Round-robin	0.2386	0.2358
Raw-score	0.0642	0.3067
Norm (max)	0.2899	0.2646
Biased RR	0.2639	0.2613
Z-score	0.2669	0.2867
Logistic	**0.3090**	**0.3393**

Multilingual IR

- Using QT approach and merging
  - Logistic regression work well (learn on CLEF 2003, eval on CLEF 2004 queries and it works well)
  - Normalization is usually better (e.,g., Z-score or divided by the max)
  - But when using the same IR system (Cond C), raw-score merging (simple) could offer an high level of performance
- For better merging method see CMU at CLEF 2005
- Berkeley at CLEF 2003
  - Multilingual with 8 languages
    QT: 0.3317   DT (into EN): 0.3401
    both DT & QT (and merging): 0.3733
- Using both QT and DT, the IR performance seems better (see CLEF 2003 multilingual (8-languages) track results)

Multilingual IR (CLEF-2003)



Conclusion

- Search engines are mostly language independent
- Monolingual
  - could be relatively simple for foreign languages close to English (Romance and Germanic family)
  - the same for Slavic family?
  - compound construction is important DE
  - more morphological analysis could clearly improved the IR performance (FI)
  - segmentation is a problem (ZH, JA)
  - no clear conclusion with KR, HU
  - some test-collections are problematic (AR in TREC 2001, RU in CLEF 2004)

Conclusion

- Bilingual / Multilingual
  - various translation tools for some pairs of language (mainly with 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)
  - we ignore a large number of languages (Africa)

The Future

- Effective user functionality
  - Effective feedback, translation, summarization
- New, more complex applications
  - CLIR factoid question
- Languages with sparse data
- Massive improvement in monolingual IR
  - Learning semantic relationships from parallel and comparble corpora
- Merging retrieval results lists form databases in multiple languages
  - Beyond shallow integration of translation tools
- More tightly integrated models for CLIR

References

Abdou, S., Savoy, J. Statistical and comparative evaluation of various indexing and search models. Proceedings AIRS-2006

Amati, G., van Rijsbergen, C.J. (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM - Transactions on Information Systems, 20, 357-389.

Brown, P., Della Pietra, S., Della Pietra, V., Lafferty, J., Mercer, R. (1993) The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), 263-311.

Braschler, M., Ripplinger, B. (2004) How effective is stemming and decompounding for German text retrieval? IR Journal, 7, 291-316.

Braschler, M. Peters, C. (2004) Cross-language evaluation forum: Objectives, results, achievements. Information Retrieval, 7(1-2), 7-31.

Braschler, M. (2004) Combination approaches for multilingual text retrieval. Information Retrieval, 7(1-2), 183-204.

Gao, J., Nie, J.-Y. (2006) A study of statistical models for query translation: Finding a good unit of translation. ACM-SIGIR'2006. Seattle (WA), 194-201.

Gale, W.A., Church, K.W. (1993) A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1), 75-102.

Grefensette, G. (Ed) (1998) Cross-language information retrieval. Kluwer.

Harman, D. (1991) How effective is suffixing? Journal of the American Society for Information Science, 42, 7-15.

References

Harman, D. (1991) How effective is suffixing? Journal of the American Society for Information Science, 42, 7-15.

Harman, D.K. (2005) Beyond English. In “TREC experiment and evaluation in information retrieval”, E.M. Voorhees, D.K. Harman (Eds), The MIT Press.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., Järvelin, K. (2004) Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000–2002. Information Retrieval, 7 (1-2), 99-119.

Hiemstra, D. (2000) Using language models for information retrieval. CTIT Ph.D. thesis.

Kraaij, W. (2004) Variations on language modeling for information retrieval. CTIT Ph.D. thesis.

Krovetz, R. (1993) Viewing morphology as an inference process. ACM-SIGIR'93. Pittsburgh (PA), 191-202.

Le Calvé A., Savoy J. (2000) Database merging strategy based on logistic regression. Information Processing & Management, 36(3), 341-359

McNamee, P., Mayfield, J. (2004) Character n-gram tokenization for European language text retrieval. IR Journal, 7(1-2), 73-97.

Nie, J.Y., Simard, M., Isabelle, P., Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. ACM-SIGIR'99, 74-81.

References

Porter, M.F. (1980) An Algorithm for suffix stripping. Program, 14, 130-137.

Savoy, J. (1993) Stemming of French words based on grammatical category. Journal of the American Society for Information Science, 44, 1-9.

Savoy J. (2004) Combining multiple strategies for effective cross-language retrieval. IR Journal, 7(1-2), 121-148.

Savoy J. (2005) Comparative study of monolingual and multilingual search models for use with Asian languages. ACM -Transaction on Asian Language Information Processing, 4(2), 163-189.

Savoy J. (2006) Light stemming approaches for the French, Portuguese, German and Hungarian languages. ACM-SIAC,1031-1035.

Sproat, R. (1992) Morphology and computation. The MIT Press.

Xu, J., Croft, B. (1998) Corpus-based stemming using cooccurrence of word variants. ACM -Transactions on Information Systems, 16, 61-81.

Xu, J., Weischedel, R., Nguen, C. (2001) Evaluating a probabilistic model for crosslingual retrieval. ACM –SIGIR-2001, New Orleans, 105-110.

Zhang, Y., Vines, P., Zobel, J. (2005) Chinese OOV translation and post-translation query expansion in Chinese-English cross-lingual information retrieval. ACM -Transactions on Asian Language Information Processing, 4 (2), 57-77