
**Multilingual Information Retrieval and Cross-Language Retrieval**

*Martin Braschler, Zürich University of Applied Sciences, Switzerland*

TrebleCLEF Summer School on
Multilingual Information Access
Hotel Santa Croce in Fossabanda, Pisa, Italy
15-19 June 2009

# Multilingual Information Retrieval & Cross-Language Retrieval

**J. Savoy**
*University of Neuchatel*
**M. Braschler**
*Zurich University of Applied Sciences*

www.unine.ch
www.init.zhaw.ch

## Outline

- **Information Retrieval**
- MLIA/CLIR motivation and evaluation campaigns
- Indexing
- Translation
- Matching
- Demo
- Lab Exercise

1

## Foreword

"It is an error to entertain any proposition with greater assurance than the proofs it is built upon will warrant"
John Locke, *Essay on Human Understanding*, 1690.

"No theory can ever be proven to be true;  it is only true until a better theory can be found"
T. L. Wilkins, *Social Deviance*, 1964

2

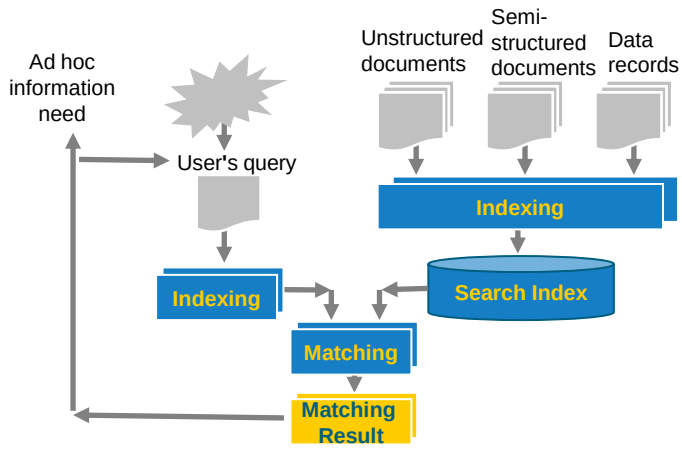
## Information Retrieval (IR)

- „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 information need)
- Result is presented in the form of a ranked list of documents (that are potentially relevant)
- Usually equalled with full text retrieval of natural-language documents. Advanced indexing and matching methods are employed

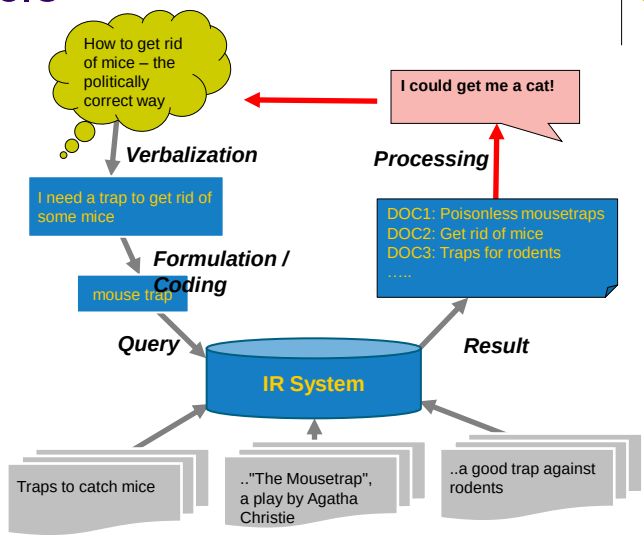
# The Retrieval Problem

- Retrieval problem: „To retrieve as much relevant information as possible while at the same minimizing the amount of irrelevant information returned“.
- Issues:
  - mismatch between document and query due to language ambiguity (synonym, homonym, paraphrasing, metaphor, word forms, 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
  - conflicting goals: maximizing relevant information vs. minimizing irrelevant information
  - relevance is subjective and context-dependent

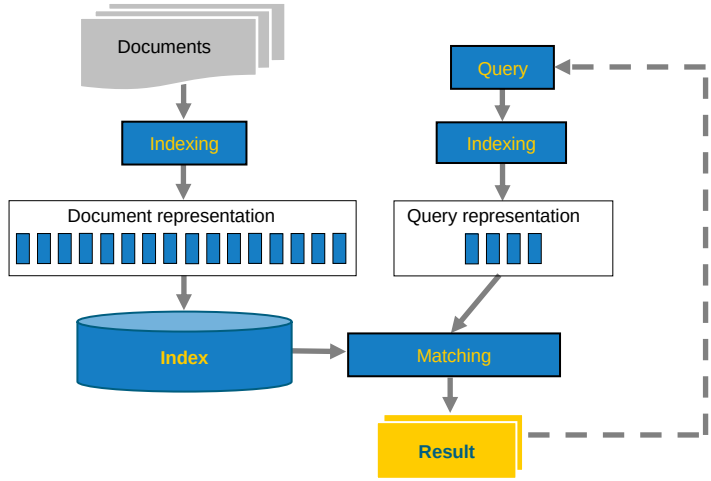
# Matching Queries & Documents



# "IR-Cycle"




# IR "Flow"



## 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]

8

## Different Degrees of MLIA/CLIR

- Monolingual retrieval in non-English languages
- Bilingual retrieval A → B
- Multilingual retrieval A → A, B, ...
- Multilingual retrieval AB → A, AB, AC, B, BC, ..

## MLIA/CLIR

- Multilingual Information Access/Multilingual Retrieval encompasses all four defintions
- Cross-Language Information Retrieval means at least a bilingual retrieval between two different languages

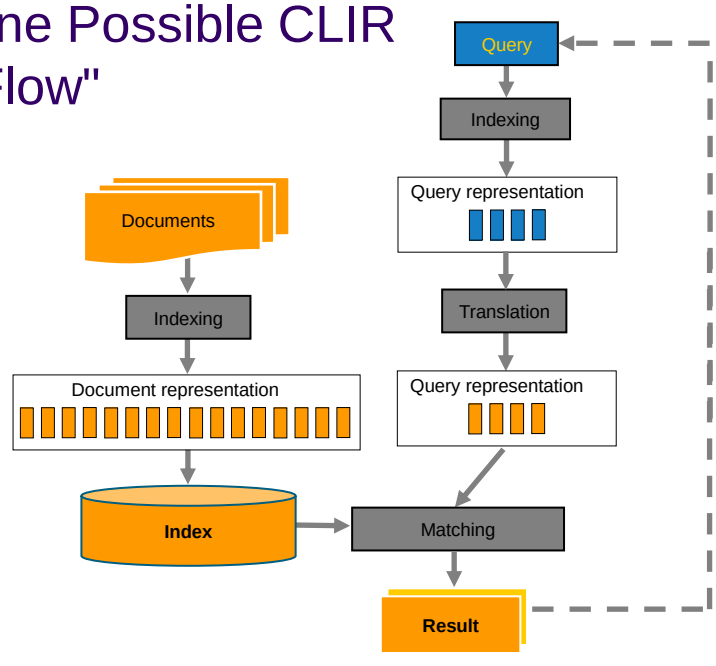
## What to Translate

- It seems evident that some form of translation is needed to bridge the language gap
- We can translate
  - The queries
  - The documents
  - Both
  - Neither (!)

# MLIA/CLIR "flow"/"structure"

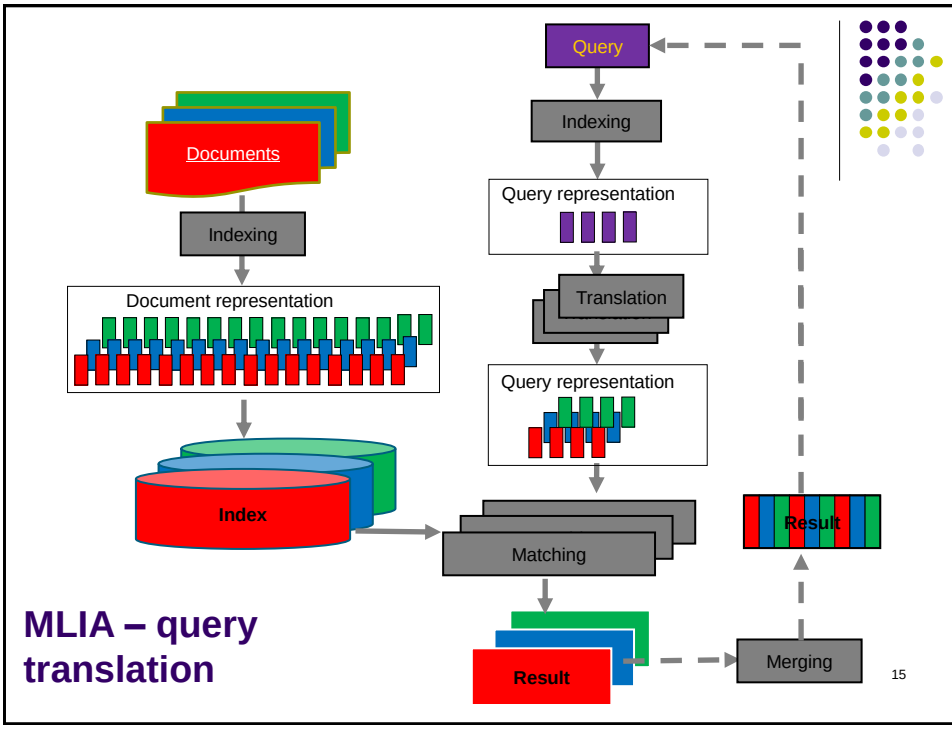
- Building a MLIA/CLIR system involves adressing different processing steps.
- We structure our discussion into the following list of steps
  - Indexing
  - Translation
  - Matching

# One Possible CLIR "Flow"



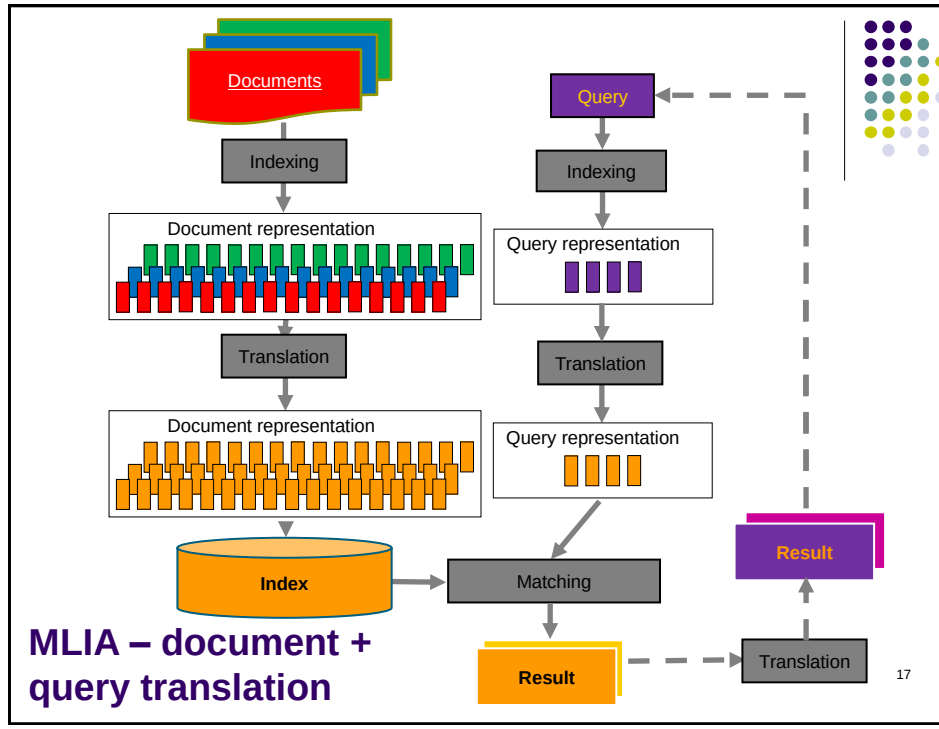
# Bilingual CLIR

- Maybe the "simplest scenario"
- We add query translation to a monolingual IR system
- How to integrate the translation step into the overall system?



# MLIA – Query Translation

- More complex setup
- A series of bilingual steps
- A merging step is needed to produce a single, integrated result



## MLIA – Document Translation

- All documents are translated into a single language
- Caveat: what happens if many query languages are possible?
- $\rightarrow$ combination with query translation, interlingua
- No need for merging step!

## Indexing

1. (Format conversion), Character conversion, Pre-processing
2. Language identification
3. (document formation)
4. Segmentation, Tokenization, Parsing
5. Feature normalization
6. Enrichment (entity recognition, ..)

## Translation

- (Automatic) Translation process
  - Problems
  1. Query translation QT
  2. Document translation DT
  3. Combined
  4. Language-independent processing
- Pre-translation Expansion
- Evaluation

## Matching

- Matching
- Weighting schemes
- Effectiveness
- Merging

## Outline

- Information Retrieval
- **MLIA/CLIR motivation and evaluation campaigns**
- Indexing
- Translation
- Matching
- Demo
- Lab Exercise

22

## 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.

  Only 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.

- see also www.omniglot.com

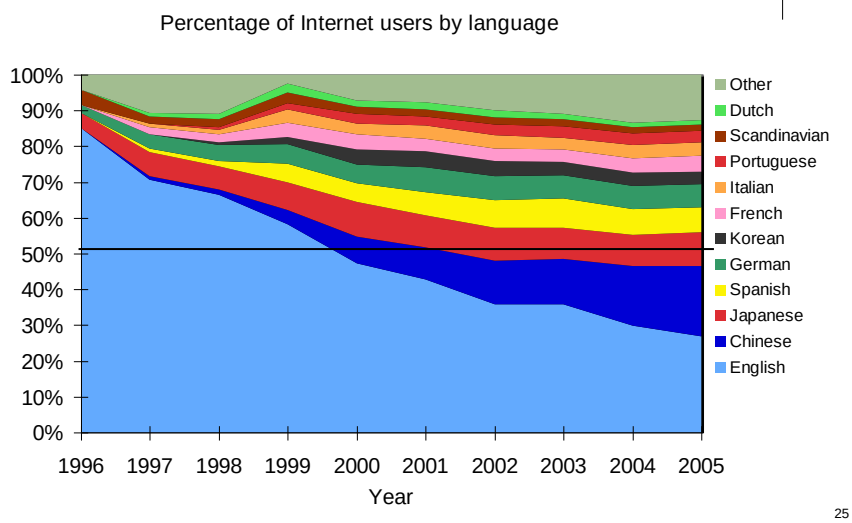
23

# 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

24

# Motivation




25

# Motivation

- 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, …

26

# Business Cases

- Bilingual / multilingual
  - 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 annoted in a different language (*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).

27

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

28

# 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
  - 2008-09: Persian
  - 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]

29

## 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 |

30

## Evaluation Campaigns

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

31

# 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

32

# 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)
  - NTCIR-7 (2009): JA, KR, ZH (traditional & simplified), IR4QA, CCLQA, MOAT, MuST, Patent translation & mining

33

## 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 |

34

## Evaluation Campaigns

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

35

## Evaluation Campaigns (FIRE 2008)

| | Hindi | Bengali | Marathi |
|---|---|---|---|
| Size MB | 718 MB | 732 MB | 487 MB |
| Docs | 95,215 | 123,047 | 99,357 |
| # token/ doc | 356 | 292 | 265 |
| # queries | 45 | 50 | 49 |
| # rel. doc./ query | 76.36 | 37.26 | 22.35 |

36

## 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*
- Pooling
  - Retrieve documents using several techniques
  - Judge top *n* documents for each technique (blind)
  - Relevant set is union
  - The result is a subset of true relevant set

37

## Precision & Recall

- Precision
  - Proportion of a retrieved set that is relevant
- Recall
  - Proportion of all relevant documents in the collection included in the retrieved set
- Precision and recall are well-defined for *sets*
- For *ranked* retrieval ?
- Often want a single-number effectiveness measure
- Average precision (AP) is widely used in IR
- Calculate by averaging precision when recall increases (at each new relevant and retrieved document)

38

## Average Precision

| Rank | System A | | System B | |
|---|---|---|---|---|
| 1 | R | 1/1 | nR | |
| 2 | R | 2/2 | R | 1/2 |
| 3 | nR | | R | 2/3 |
| … | nR | | nR | |
| 35 | nR | | R | 3/35 |
| … | nR | | nR | |
| 108 | R | 3/108 | nR | |
| | AP = | 0.6759 | AP = | 0.4175 |
| | | | | -38.2% |

39

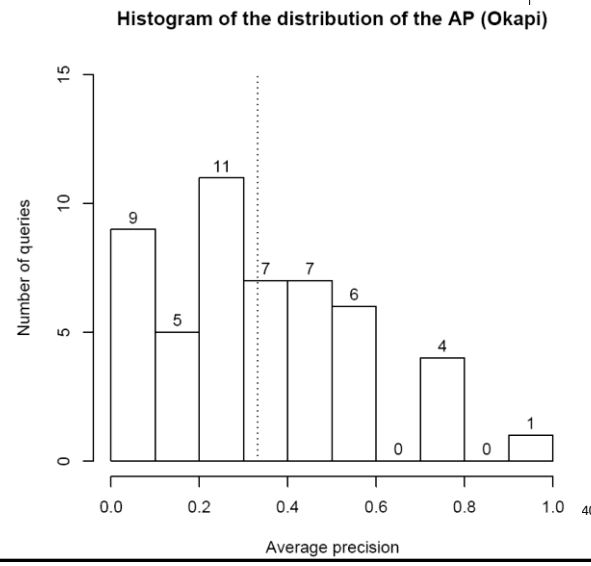
# Mean Average Precision (MAP)

A single value
MAP: 0.3321
or an histogram?

Here, for one query, the perfect answer

For 9 queries, Okapi "fails"

(ZH, NTCIR-5, indexing unigram & bigram)




40

# Significance Tests

- Are observed differences statistically different?
- Generally can't make assumptions about underlying distribution (non-parametric or parametric test)
  - Most significance tests *do* make such assumptions
- Various statistical tests are possible
  - Sign test
  - Wilcoxon signed-ranks test
  - *t*-test
  - Boostrap test
- Are observed differences detectable by users?

41

## MRR

- Mean Reciprocal Rank (of the first good answer)
- MRR (= 1 / rank) penalizes a false answer in the first position
- Precision-oriented measure
- From our previous example
  System A: MRR = 1
  System B: MRR = 0.5

| Rank | MRR |
|---|---|
| 1 | 1.000 |
| 2 | 0.500 |
| 3 | 0.333 |
| 4 | 0.250 |
| 5 | 0.200 |
| 6 | 0.167 |
| 7 | 0.143 |
| 8 | 0.125 |
| 9 | 0.111 |
| 10 | 0.100 |

42

## Outline

- Information Retrieval
- MLIA/CLIR motivation and evaluation campaigns
- **Indexing**
- Translation
- Matching
- Demo
- Lab Exercise

43

# Indexing Step 1

- (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

<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>

45

## 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

46

## Even English is not Just English

- Historical variations in English
  The need to use the same language for the query formulation!
  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 eorthan swa swa on heofonum.

47

## 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

48

## 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)

49

## Monolingual IR

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

50

## Indexing Step 2

- 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

- 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`, ?

52

# 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

53

## Indexing Step 3

- What is the granularity of retrieved items?
  - Entire document
  - Sub-document (chapter, paragraph, passage, sentence)
  - Super-document (aggregation of documents, linked documents, folders)

→ Will not be discussed further (see, e.g., XML IR)

## Indexing Step 4

- The document is split into "valid" tokens
- The tokens are suitable to form the index structure
- "Undesirable" tokens are eliminated
  - non-content bearing tokens
  - special characters
  - (numbers, date)
  - very short or very long tokens, ...

## Monolingual IR (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 …

  Database system
  data base system
  data-base system (hyphen ?)

56

## 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].
  - → see step 5

57

## 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))

58

## 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)

59

# Monolingual IR

- Language independent approach
  *n*-gram indexing [McNamee & Mayfield 2004], [McNamee 2008]
  - 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

62

# Monolingual IR

A Korean compound term

정보검색시스템

words

정보검색시스템

Bigrams

정보 보검 검색 색시 시스 스템

Decompounded (HAM)

정보 검색 시스템

63

# 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 [Savoy 2005]

| MAP / ZH (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 |

64

# Monolingual IR

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

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

Baseline in bold, difference statistically significant underlined

| MAP / 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 |

65

## Monolingual IR

- Stopword lists
  - 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)

66

## Monolingual IR

- 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 (TLR, Thomson Legal & Regulatory, now Thomson Reuters) uses one term ("the")

67

## Monolingual IR

Evaluation CLEF 2001 to CLEF 2006 (*Los Angeles Times* (1994) & *Glasgow Herald* (1995)), for 169,477 documents and 284 TD queries)

| **MAP** | SMART (571 words) | Short (9 words) | None |
|---|---|---|---|
| Okapi | 0.4516 | 0.4402 | 0.3839 |
| DFR-I($n_e$)B2 | **0.4702** | **0.4743** | **0.4737** |
| DFR-PL2 | 0.4468 | 0.4463 | 0.3159 |
| DFR-PB2 | 0.4390 | 0.3258 | 0.0287 |
| *tf idf* | 0.2742 | 0.2535 | 0.2293 |

Underlined: significant difference with SMART



## Monolingual IR

Evaluation CLEF 2001 to CLEF 2006 (*Le Monde* (94-95) & *SDA* (94-95)), for 177,452 documents and 299 TD queries

| **MAP** | Long Stoplist (464 words) | Short (20 words) | None |
|---|---|---|---|
| Okapi | 0.4321 | 0.4286 | 0.2457 |
| DFR-I($n_e$)B2 | **0.4499** | **0.4490** | **0.4467** |
| DFR-PL2 | 0.4247 | 0.4216 | 0.3080 |
| DFR-PB2 | 0.4167 | 0.4172 | 0.0469 |
| *tf idf* | 0.2867 | 0.2758 | 0.2436 |

Underlined: significant difference with "Long StopList"

## Monolingual IR

- 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.

70

## Monolingual IR

- Problem with the Okapi formulation
  the *idf* weight (probabilistic interpretation)
  with *n* the number of document
  $df_j$ the number of documents indexed by term $t_j$

$$idf_j = \log\left(\frac{n - df_j + 0.5}{df_j + 0.5}\right)$$

  and if $df_j > n/2$, the *idf* value becomes negative!
- A solution could be

$$idf_j = \log\left(1 + \frac{n - df_j + 0.5}{df_j + 0.5}\right)$$

71

## Monolingual IR

- Problem with DFR paradigm
  Some implementations (e.g., PL2, PB2) compare the expected number of occurrences with the observed frequency.

  Example: the expected number of “the” in an article owning 3,409 terms is largely greater than the observed number (1 in this case, sport results).
  A real *Divergence from Randomness*!
  The corresponding article will be the first retrieved item (for all queries having a “the”)

72

## Indexing Step 5

- 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., "database" or "data base")
  - lexical variants (e.g., "analyzing", "analysis")
  - equivalent terms that are synonymous in meaning (e.g., "film", "movie")

## 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)

74

## 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)
    Mona Lisa ↔ La Joconde ↔ La Gioconda
  - Specialized thesauri are useful (MultiMatch project)
    Unified List of Artist Names
    Arts and Architectures Thesaurus
    Thesaurus of Geographic Names

75

## Monolingual IR

- Normalization / Proper nouns
  - 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.).
  - DE: different (and contradictory) spelling reforms.
  - Think about SMS language (BTW, 4Y, P2P, …)

76

## Monolingual IR (Stemming)

- Stemming (words & rules)
  - Inflectional (*light*)
    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 (stem + suffix = word)
    forming new words (changing POS)
    ‘-ably’, ‘-ment ’, ‘-ship’
    admit → {admission, admittance, admittedly}

77

## 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]

78

## 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”

79

## 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"

80

## Monolingual IR

- Various algorithmic stemmers suggested
  - do nothing
  - inflectionnal stemmer (S-stemmer)
  - derivational stemmers (Porter, Lovins, SMART)
- Under "Lemma", result of the morphological analysis (see Robust track at CLEF-2008)
  - We may also consider POS (Part-of-Speech)
  - Synset returned by WordNet

81

## Monolingual IR

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

| | None | S-stem | Porter | Lovins | SMART | Lemma |
|---|---|---|---|---|---|---|
| Okapi | **0.4345** | 0.4648† | 0.4706† | 0.4560 ‡ | 0.4755† | 0.4663† |
| PL2 | 0.4251 | 0.4553† | 0.4604† | 0.4499†‡ | 0.4634† | 0.4608† |
| $I(n_e)C2$ | 0.4329 | **0.4658**† | **0.4721**† | **0.4565** ‡ | **0.4783**† | **0.4671**† |
| LM | 0.4240 | 0.4493† | 0.4555† | 0.4389 ‡ | 0.4568† | 0.4444† |
| *tf idf* | 0.2669 | 0.2811† | 0.2839† | 0.2650 ‡ | 0.2860† | 0.2778† |
| Average | 0.4291 | 0.4588 | 0.4647 | 0.4503 | 0.4685 | 0.4597 |
| %change | | +6.9% | +8.3% | +4.9% | +9.2% | +7.1% |

underlined: significant with the best (column)
† with "None"
‡ with "SMART"

82

## Monolingual IR

- 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).

83

## Monolingual IR

| | Lemma | Lemma & POS | Lemma & Synset | Lemma &POS+Synset |
|---|---|---|---|---|
| Okapi | 0.4663 | 0.4720† | 0.4395† | 0.4482† |
| PL2 | 0.4608 | 0.4634 | 0.4365† | 0.4433† |
| I($n_e$)C2 | **0.4671** | **0.4740†** | **0.4665** | **0.4705** |
| LM | 0.4444 | 0.4562† | 0.4342† | 0.4458 |
| *tf idf* | 0.2778 | 0.2879† | 0.2834 | 0.2888† |
| Average | 0.4597 | 0.4664 | 0.4442 | 0.4520 |
| %change | | +1.5% | -3.4% | -1.7% |

underlined: significant with the best (column)
† with "Lemma"

84

## Monolingual IR

- Topic #217 (“AIDS in Africa”)
  - AP = 0.1944 under "Lemma"
  - AP = 0.5526 with lemma & POS
  - The term “AIDS” into “aid” but tags as proper name (NNP).
- And for other languages (having a more complex morphology)?
  - French (and other Latin languages)
  - German
  - Slavic languages (Bulgarian, Czech, Russian)
  - Hungarian (Finnish)

85

# Monolingual IR (French)

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

86

# Monolingual IR (French)

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 |

Underlined: difference statistically significant with bold

87

# Monolingual IR (French)

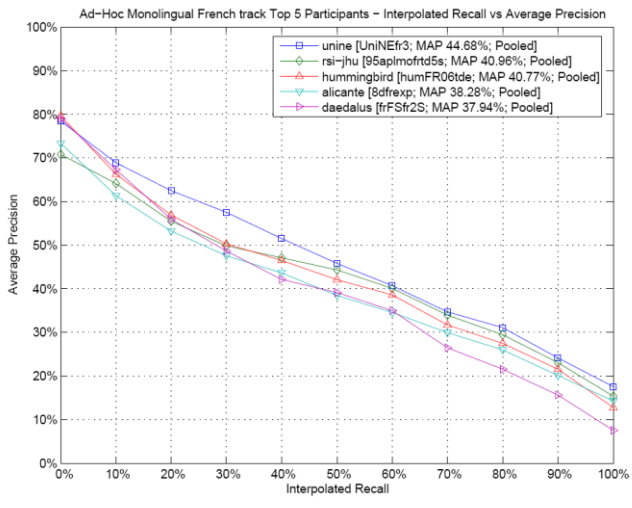
Based on CLEF-2005 corpus, T queries

| 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 |

Underlined: difference statistically significant with bold

88

# Monolingual IR (CLEF 2006)




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

89

# 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")
- Example with Italian
  Plural forms for nouns
  -e → -e ("cane", "cani")
  -a → -e ("rosa", "rose"), …
  Feminine forms -o → -a ("amico" → "amica")

90

# 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")
- 3 genders x 2 numbers x 4 cases = 24 possibilities!
- Compound construction
  ("Lebensversicherungsgesellschaftsangestellter"
  = life + insurance + company + employee)

91

## Monolingual IR (Stemming)

- Bulgarian (9M, Southern Slavic language, CLEF 2005-07)
  - Cyrillic
  - No grammatical cases
  - Definite article
- Czech (11M, Western Slavic language, CLEF 2007)
  - Latin
  - Seven grammatical cases
  - Suffixes also for names
- Russian (165M, Eastern Slavic language, CLEF 2002-08)
  - Cyrillic
  - Six cases

92

## Monolingual IR (Bulgarian)

- Inflections (definite article (“the”) and plural form)
  - море
    sea
  - морето
    the sea
  - морета
    seas
  - моретата
    the seas
- Really unknown pattern in the English language?
  Not really (from Arabic language) “alchemy”, “algebra”

93

# Monolingual IR (Bulgarian)

- Inflectional (gender, number, definite article)
  - слаб — weak (masc, sing)
    слаба (femi, sing)
    слабата (femi, sing, the)
- Inflectional & derivationals
  - българ — « stem »
    България — Bulgaria (noun)
    българин — Bulgarians (noun, masc, sing)
    българка — Bulgarians (noun, femi, sing)
    българи — Bulgarians (noun, masc, plur)
    български — Bulgarian (adj, m sing or m/f/n plur)
    българска — Bulgarian (adj, femi, sing)
    българските — the Bulgarians (adj, masc, plur)

94

# Monolingual IR (Bulgarian)

- Mutation: –я–
  - бял → белота (white → whiteness)
  - грях → грехове (sin → sins)
- Elision of vowel: –е– or –ъ–
  - орел → орли (eagle → eagles)
  - топъл → топла (warm, masc → femi)
- Palatalisation: к, г, х → ч, ж, ш
  - око → очи (eye → eyes)
  - бог → боже (God, nom → voc)
- Other: к, г, х → ц, э, с
  - вълк → вълци (wolf → wolves)
    геро → геройят (hero → heros)

95

# Monolingual IR (Stemming)

Stemming strategies, Bulgarian langauge
Based on CLEF-2006-07 corpus, 99 queries

| BU (TD) | none | UniNE | Nakov' |
|---|---|---|---|
| Okapi | 0.2115 | 0.2805 | 0.2642 |
| *tf·idf* | 0.1697 | 0.1937 | 0.2013 |

Stopword list

| BU (TD) | none | UniNE | BTB |
|---|---|---|---|
| Okapi | 0.2739 | 0.2805 | 0.2796 |
| *tf·idf* | 0.1928 | 0.1937 | 0.1930 |

96

# Monolingual IR (Czech)

- Latin alphabet (with diacritics)
- Seven grammatical cases

| case / gendre | nominative | dative singulier | dative plural |
|---|---|---|---|
| Masculine (sir) | pán | pánovi | pánům |
| Feminine (woman) | žena | ženě | ženám |
| Neutre (young) | mladé | mladému | mladým |

97

# Monolingual IR (Czech)

- Even for names

| Case | Paris | Praha | Francie | Ann |
|---|---|---|---|---|
| nominative | Pařiž | Praha | Francie | Anna |
| genitive | Pařiže | Prahy | Francie | Anny |
| dative | Pařiži | Praze | Franciі | Anně |
| accusative | Pařiž | Prahu | Franciі | Annu |
| vocative | Pařiži | Praho | Francie | Anno |
| locative | Pařiži | Praze | Franciі | Anně |
| instrumental | Pařiží | Prahou | Francií | Annou |

98

# Monolingual IR (Czech)

- Consonant softening
  - matka → matčin (mother → mother's)
  - drahý → drazí (dear, nominative sing → plur)
  - mokrý → mokří (wet, nominative sing → plur)
  - český → čeští (Czech, adje nominative sing → plur)
- Fleeting – e –
  - zámek → zámkem (castel, nominative → instrumental)
  - otec → otcův (father → father's)
- ů → o
  - stůl → stoly (table → tables)
- Derivationals
  - klavírista (piano → pianist, man)
  - klavíristka (piano → pianist, woman)
  - Židovka (Jewish woman)

99

# Monolingual IR (Stemming)

Stemming strategies, Czech language
Based on CLEF-2008 corpus, 50 queries

| CZ (T) | none | UniNE | Aggr. |
|---|---|---|---|
| Okapi | 0.2040 | 0.2990 | 0.3065 |
| *tf·idf* | 0.1357 | 0.2040 | 0.2095 |

Underlined: difference statistically significant with "none"
With and without stopword list
performance differences around 1%

100

# Monolingual IR (Russian)

- Cyrillic alphabet
- Six grammatical cases

| case / gendre | nominative | dative singular | dative plural |
|---|---|---|---|
| Masc. hard (city) | город | городу | городам |
| Masc. soft (husband) | муж | мужу | мужьям |
| Feminine (hand) | рука | руке | рукам |

101

# Monolingual IR (Stemming)

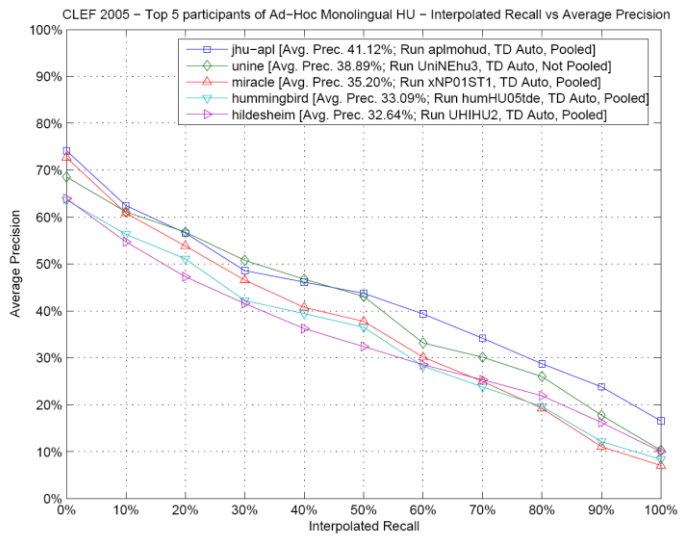
Finno-Hungarian family owns numerous cases (18 in HU, 15 FI)

| | |
|---|---|
| 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 (“internetfüggők”, “rakkauskirje”)

102

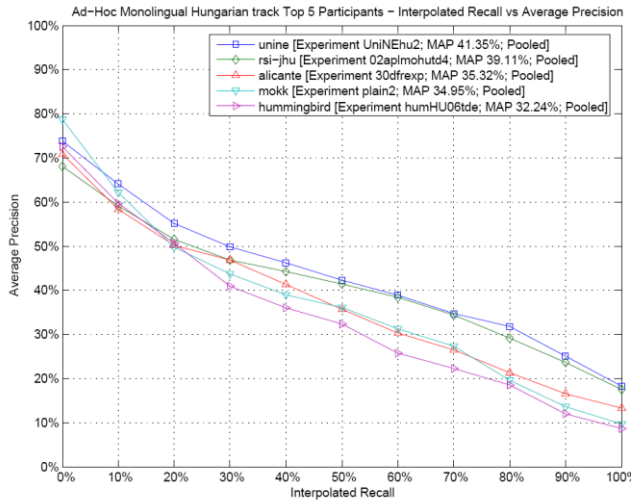
# Monolingual IR (CLEF 2005)




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

103

# Monolingual IR (CLEF 2006)




- But it change over time

104

# 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

105

# 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 = **kitab**

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

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

106

# 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], [McNamee & al. 2009]
- As a variant, use trunc(*n*), extract only the first *n* characters of each word
  trunc(5): "bookshop" → "books"

107

## Monolingual IR (Stemming)

- Evaluations & experiments in CLEF / NTCIR proceedings
- Main trends (MAP)
  - Stemming > none
  - Differences between stemmers could be stat. significant
  - Light 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
- Various applications (IR, Summarization, NLP)
- Does the user see (need to see) the stemmed form?

108

## Monolingual IR (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

109

# Monolingual IR (Lexical Links)

- Lexical relationships between languages
  - “paprika”, “goulash”, “saber” from HU
  - “robot” from CZ
- But the dominant language tends to impose its new words
  - modern, interview, sport, jury, pedigree, computer, internet, CD, DVD, cassette, snob, pub, microwave, …
- Examples
  - disc (EN) → “disk” (e.g., CZ)
    → “disc” (using the Latin letters)
    → “диск” (in Russian, Cyrillic letters)
  - Renault (EN) → “Renault” (e.g., CZ)
    → “Рено” (in Russian, Cyrillic letters)
  - CLEF topic “(Best Picture) Oscar” vs. “Oskar”

110

# Monolingual IR

- Word (stem) or *n*-gram?
- *n*-gram [McNamee, 2008] [McNamee et al., 2009]
  - Effective for Far-East languages (ZH, JA, KR)
  - Language-independant approach
  - Could be used for European languages (with *n* = 3 to 5)
    Example: "Pesticides in Baby Food"
    5-gram "pesti estic stici ticid icide cides baby food"
  - We can ignore the stemming and stopword list problem
  - Useful with noisy data (OCR ?)
  - Always effective ?

111

## Monolingual IR

- *n*-gram or stem?
- Topic #306 (“ETA Activities in France” or “ETA-tevékenységek Franciaországban” (HU)), 6 relevant items.
  AP = 0.0101, 4-gram (Okapi)
  AP = 0.5807, word (& decompounding) (Okapi)
  Query = {“eta”, “tevekenyseg” (activity), “franci” (French), “franciaorszag” (France), “franci” (French) and “orszag” (country)}
- The problem with 4-gram?
  *Multiple* matches on “Franciaországban”, “tevékenységek”.
  Retrieved many non-relevant documents (“France Télécom” or “Jacques Chirac”) but not with right actor (ETA in this case)

112

## Monolingual IR

- *n*-gram or stem?
- Topic #315 (“Doping in Sports” or “Doppingolás a sportban”) (HU), 73 relevant items.
  AP = 0.6713 , 4-gram (Okapi)
  AP = 0.289, word (& decompounding) (Okapi)
  Query = {“doppingol”, “spor”}.
- The advantage with 4-gram?
  *Multiple* matches (but not too many!) on the "doping" concept which clearly boosted the number of relevant articles
- After normalizing the surface forms, we need to store them in a effective and efficient manner!

113

## 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” = “радио” (radio) + “апарат” (receiver)
  - FI: “työviikko” = “työ” (work) + “viikko” (week)
  - HU: “hétvégé” = “hét” (week / seven) + “vég” (end)
- Compound may have an impact on retrieval effectiveness

114

## 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)

115

# (German) Decompounding

- How to automatically decompound German words (e.g., “Atomtests,” “Wintersport”, “Bundesrat”)?
- The composition can be done without glue
  → "Atom" + "Tests" = “Atomtests”
  or with (e.g., "s", "es" or "ens" in DE, only "s" in SW)
  → "Bund" + es + "Rat" = “Bundesrat”
- Each word is view as an instance of the pattern
  Pg = « Head + glue +Tail » or « H-g-T »
- Use impossible or infrequent trigrams in the corresponding language (e.g. the sequence “fff” is impossible in German, thus “Schifffahrt” is a compound built as “Schiff+fahrt” )

116

# (German) Decompounding

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

| | | | |
|---|---|---|---|
| computer | **2452** | port | **1091** |
| computers | **79** | ports | **2** |
| sicherheit | **6583** | sport | **1483** |
| sicher | **4522** | winter | **1643** |
| bank | **9657** | winters | **148** |
| bund | **7032** | wintersport | **44** |
| bundes | **2884** | wintersports | **2** |
| bundesbank | **1453** | | |
| präsident | **24041** | | |

117

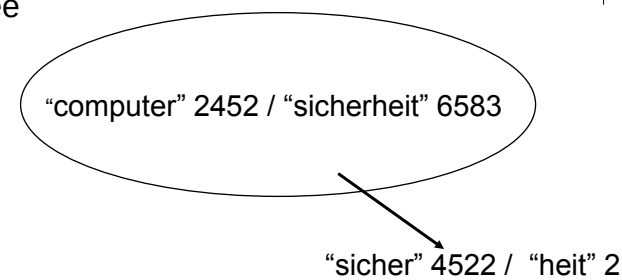
# Decompounding

- Algorithm: given a min length >*k* (*k*=3), we start at the end-*k*-1 and try to decompound the input word according to the pattern Pg. The decompound succeed if both the H and T part is in the words list.
- Example: with “computersicherheit”, we first found T=“heit”, g=“”, and H=“computersicher”. However, H does not appear, thus fails. Then find T=“icherheit,” and H=“computers”, T does not appear; fails.
  We find T=“sicherheit,” H=“computer,” and g=“s”, OK.
- We form the root of the decompounding tree with (“computer” 2452, “sicherheit” 6583). Recursively, we try to decompound both the H and T parts.

118

# Decompounding

The final tree




One possibility is to consider the occurrence frequencies of both the compound (e.g, 6583 for “sicherheit” and its composite parts (4522+2)). We may choose to consider only the most frequent node (“sicherheit” in this case).

119

## Decompounding

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)

120

## Indexing Step 6

- 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, ...)
  - ...

## Indexing: Result

- The result from the preceeding six indexing step is a "stream" of features.
- In monolingual IR, these features are used as the basis for matching
- In CLIR/MLIA, these features are also the input for the translation step (document features or query features)
- Features are treated as "bag of words" (or, more precisely, "bag of features")

## Indexing: Bag of Words Assumption

- Bag of words assumption
  - The dog bites the man → bite, dog, man
  - The man bites the dog → bite, dog, man
- But many languages add suffixes to denote the grammatical cases (subject, direct object, etc.)
  - Canis mordet hominem
  - Canem mordet homo

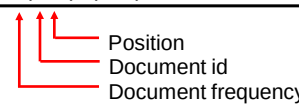
# Inverted Index

- Credits for this example to H.-P. Frei
- After indexing, we create an inverted index
- Access is by looking up features, and processing the associated lists of documents

| Doc. ids | Text |
|---|---|
| 1 | Pease porridge hot, pease porridge cold |
| 2 | Pease porridge in the pot |
| 3 | Nine days old |
| 4 | Some like it hot, some like it cold |
| 5 | Some like it in the pot |
| 6 | Nine days old |

# Inverted Index

| Feature # | Feature | df, document ids, positions |
|---|---|---|
| 1 | cold | 2; (1, 6), (4, 8) |
| 2 | days | 2; (3, 2), (6, 2) |
| 3 | hot | 2; (1, 3), (4, 4) |
| 4 | in | 2; (2, 3), (5, 4) |
| 5 | it | 2; (4, 3,7), (5, 3) |
| 6 | like | 2; (4, 2,6), (5, 2) |
| 7 | nine | 2; (3, 1), (6, 1) |
| 8 | old | 2; (3, 3), (6, 3) |
| 9 | pease | 2; (1, 1,4), (2, 1) |
| 10 | porridge | 2; (1, 2,5), (2, 2) |
| 11 | pot | 2; (2, 5), (5, 6) |
| 12 | some | 2; (4, 1,5), (5, 1) |
| 13 | the | 2; (2, 4), (5, 5) |



## Inverted Index: Consequences

- Access to features is very fast (hash table lookup), scales very well
- Easy to compute various statistics (*tf*, *idf*, collection frequency)
- However, access is only efficient if performing exact matches of features
- No efficient handling of
  - wildcards
  - substring searches
  - "complex features" such as phrases...
- → The right segmentation and normalization is crucial!

## Outline

- Information Retrieval
- MLIA/CLIR motivation and evaluation campaigns
- Indexing
- **Translation**
- Matching
- Demo
- Lab Exercise

127

# Translation

Difficult problem, even for humans

- *Rome, Italy*
  "Please dial 7 to retrieve your auto from the garbage"
- *India*
  "Children soup"
- *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"

C. Crocker: *Løst in Tränšlatioπ. Misadventures in English Abroad.* O'Mara Books, London, 2006

128

# Translation

Difficult problem, even for humans

- *Chelsea, London*
  "Plat du jour: changed each day"
- *Pizza Restaurant, London*
  "Open 24 hours except 2 a.m. – 8.a.m."
- *A Mexican bar*
  "Sorry, we're open!"

C. Crocker: *Løst in Tränšlatioπ. Misadventures in English Abroad.* O'Mara Books, London, 2006

129

## Translation Problem

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

130

## Translation Problem

- (Manual) Translation possibilities
- 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”)

131

# Translation Ambiguity

- "post"

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

- "temps" (FR) → time, weather, tense
- "light" (EN → FR), POS may help

| | |
|---|---|
| noun | "lumière" |
| adjective | "clair", "léger" |

132

# Translation

- 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)
- In a bilingual country, translation is not so expensive
  Is Canada a bilingual country?
- More complex in real multilingual organizations
  In EU, with 23 languages, we need to provide $(23 \cdot 22)/2 = 253$ language pairs!

133

# Automatic Translation (Example)

- "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 "

134

# Automatic Translation

- What do we need to translate?
- Topic translation (QT)
  - less expensive
- Documents translation (DT)
  - done before the search
- Mixed query and documents translation
  - could be very effective

135

## 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 name in Italian"

136

## 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., “car woman bag and man walking in a street"
    or difficult to understand/classify
    “plate orange” 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., TEL)

137

## Automatic Translation

- Automatic translation will add ambiguity in the IR process
  - Multiple translations of each word
  - Use translation probabilities (how?)
    (MT system is a black box)
  - 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

138

## Automatic Translation

- Example of (language) resources
  - CIA facts book (names)
    see http://www.cia.gov
  - Web
    ZH → EN: the name in Chinese follows the name written in English
  - Wikipedia
  - Parallel corpora
    UN web site
    EU (& Official Journal)
  - Specialized thesauri

139

# Automatic Translation

- Specialized thesaurus
  - GIRT (German Indexing and Retrieval Test database) CLEF 2001 – 2008
  - Available DE ↔ EN
  - Example of an entry

```
<entry>
<german>            Volksabstimmung
<german-caps>       VOLKSABSTIMMUNG
<broader-term>      direkte Demokratie
<narrower-term>     Volksbegehren
<narrower-term>     Volksentscheid
<english-translation>  plebiscite
</entry>
```

140

# 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)

141

# 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
  - see project mboi at rali.iro.umontreal.ca/
  - MOSES statistical machine translation model www.statmt.org/moses/
  - Statistical translation methods tend to dominate the field
- How can we improve the translation process?

142

# Translation Strategies

- Example EN → FR (idiomatic)


143

# Translation Strategies

- Example EN → IT


144

# 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]
- Other approaches to improve the translation?

145

# 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"

146

# Pre-Translation Expansion

- Useful additional terms could be morphological related terms (British, Britain, UK)
- Two stages
  - Select the right terms to be added
  - Weight these additional terms
- Usually using the same formula (e.g., Rocchio)

$$Q_{i+1} = \alpha \cdot Q_i + \beta \cdot \frac{1}{r} \cdot \sum_{k=1}^{r} d_k - \gamma \cdot \frac{1}{nr} \cdot \sum_{k=1}^{nr} d_k$$

$$Q_{i+1} = \alpha \cdot Q_i + \beta \cdot \sum_{k=1}^{r} d_k - \gamma \cdot \sum_{k=1}^{nr} d_k$$

147

## Pre-Translation Expansion

Traditional problems [Peat & Willett, 1991]

- Original query must return reasonable retrieval results
  (we need to find relevant items in the top of the results list)
- Peat & Willett found that most query terms have a greater occurrence frequency than to do other terms.
- Query expansion approaches based on term co-occurrence data will include additional terms that also have a greater occurrence frequency in the documents.
- In such cases, these additional search terms will not prove effective in discriminating between relevant and non-relevant documents.
  The final effect on retrieval performance could be negative.

148

## Pre-Translation Expansion

Additional problems

- The search system provides indexing terms
  (not surface word or stem)
  E.g., from "Chinese currency devaluation"
  we have "chines currenc devalu" (Porter)
  The translation step cannot use such "stems"
  Could be useful to consider applying a light stemmer!
- You need not only the target corpus (written in the target language) but also a similar corpus written in the query language.

149

## 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;

150

## 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
    as well as more "controlled" source (e.g. Wikipedia)
    specialized thesauri

151

## 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

152

## Translation Strategies

- “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

153

# Translation Strategies

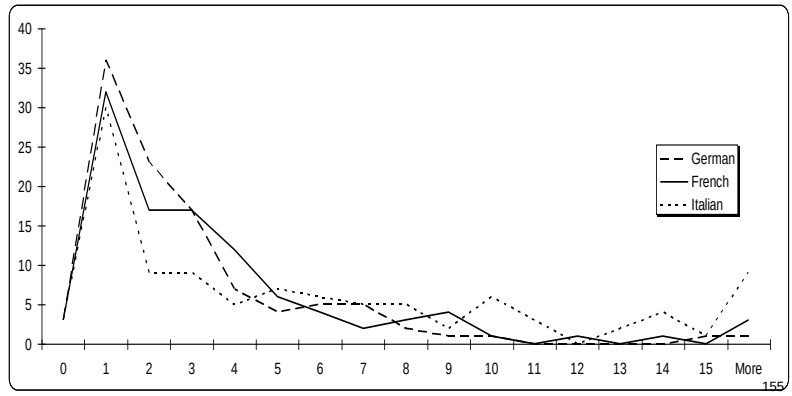
Example of phrases

- "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" (power in the political sense)
- "Mad Cow Dease "
  - in FR, instead of "maladie de la vache folle"
    "fou vache malade" (illness vs. ill)
    the stemmer does not always conflate under the same root

154

# Translation

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




155

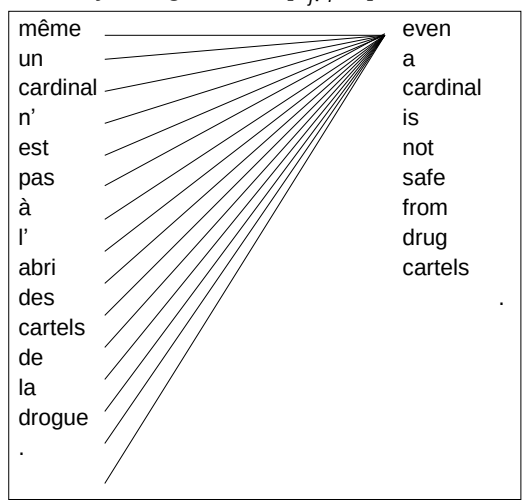
# Translation Strategies

- 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[$e_j|f_i$ , $A$]

156

# Translation Strategies

Initial probability assignment P[$e_j|f_i$, $A$]

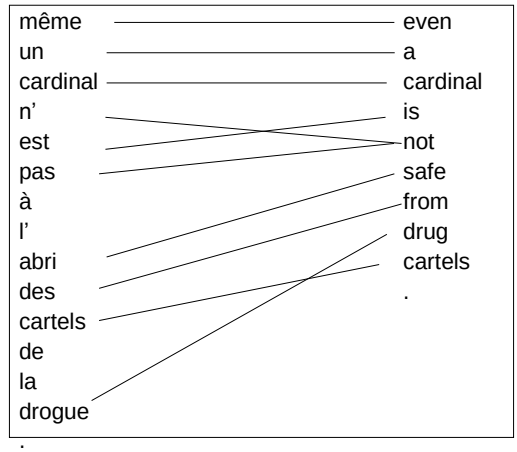



Slides from J.Y. Nie

157

# Translation Strategies

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






# Translation Strategies

With parallel corpora [Gale & Church 1991]

- Example with the mboi system (rali.iro.umontreal.ca/mboi)
- From "database system"
  in French
  "(données^0.29472154 base^0.20642714
  banque^0.037418656")
  (correct translation "système de bases de données")

# Translation

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

| Translation | Query | AP |
|---|---|---|
| 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 |

160

# 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, $\alpha$=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 |

161

# 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%) |

162

# 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 |

163

# Translation

On a large query set (299 CLEF 2001-06, French corpus)
Original query written in French (Title-only) [Savoy & Dolamic 2009]
Automatic translation using Google (May 2007)

| | **MRR** | | **MAP** | |
|---|---|---|---|---|
| | Mono | From EN | Mono | From EN |
| Okapi | **0.6631** | **0.5817** | **0.4008** | **0.3408** |
| LM | 0.5948 | 0.5093 | 0.3647 | 0.3085 |
| *tf idf* | 0.5072 | 0.3895 | 0.2591 | 0.2091 |

164

# Translation

On a large query set (284 CLEF 2001-06, English corpus)
Original query written in English (Title-only) [Dolamic & Savoy 2009]
Statistical significant difference (*)

| | **MRR** | **MAP** |
|---|---|---|
| | Mono | Mono |
| I(ne)C2 | **0.6614** | **0.4053** |
| Okapi | 0.6656 | 0.4044 |
| LM | 0.6086* | 0.3708* |
| *tf idf* | 0.4453* | 0.2392* |

165

# Translation

Original query written in English (284 T-only) [Dolamic 2009]
Automatic translation done by Google (May 2007)
Statistical significant difference (*)

| **MAP** | Mono | From ZH | From DE | From FR | From SP |
|---|---|---|---|---|---|
| I(ne)C2 | **0.4053** | *0.3340** | 0.3618* | *0.3719** | 0.3741* |
| Okapi | **0.4044** | 0.3327* | *0.3625** | 0.3692* | *0.3752** |
| LM | **0.3708** | 0.3019* | 0.3305* | 0.3400* | 0.3426* |
| *tf idf* | **0.2392** | 0.1920* | 0.2266* | 0.2294* | 0.2256* |
| *diff* | | -18.2% | -9.3% | -7.3% | -7.1% |

166

# Translation

Original query written in English (284 T-only) [Dolamic, 2009]
Automatic translation done by Yahoo (may 2007)
Statistical significant difference (*)

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

167

## 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%)

168

## Translation

Where are the real translation problems?

For Google MT system

| **Source** | ZH | DE | FR | SP |
|---|---|---|---|---|
| name | 21 | 2 | 1 | 2 |
| polysemy | 16 | 4 | 11 | 11 |
| morphology | 2 | 2 | 1 | 2 |
| compound | 0 | 4 | 0 | 1 |
| other | 0 | 0 | 2 | 0 |

169

# Translation Pivot Language

On a large query set (299 CLEF 2001-06, French corpus)
Original query written in French (Title-only) [Savoy & Dolamic 2009]
Query language is German

| **MRR** | Mono | From EN | From DE | From DE-EN |
|---|---|---|---|---|
| Okapi | **0.6631** | **0.5817** | **0.4631** | **0.5273** |
| Diff. | | -12.3% | -30.2% | -20.5% |

170

# Translation Pivot Language

Why?

- Better resources done for translations from/to English
- Compound construction in German
  Example:
  "Robbenjagd" = "Robben"(seals) + "Jagd" (hunting))
  correctly translated into English ("Seal hunting")
  not into French ("Robbenjagd").

171

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

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

172

# 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])

173

## 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 $C_l$). We may estimate as:

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

174

## 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

175

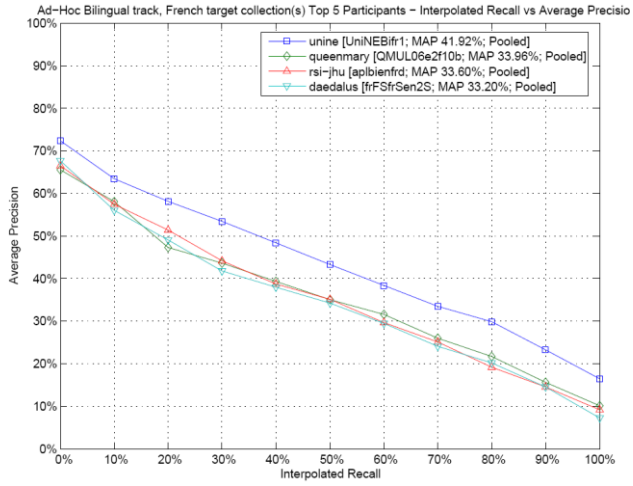
# Towards Better Translation

- Using the context
  - “temps” (FR) → time, weather, tense
  - “vol” (FR) → flight, theft, flock
  - “temps de vol” → time of flight
- Using the POS will help (EN → FR)
  - “light” noun → “lumière”
  - Adjective → “clair”, “léger”
- Domain-specific will help
  (only one meaning = one translation?

  Window (in CS) → OS?, windowing system, how to open a window in Java?, windows and UI?

176

# CLIR (CLEF-2006 X → FR)

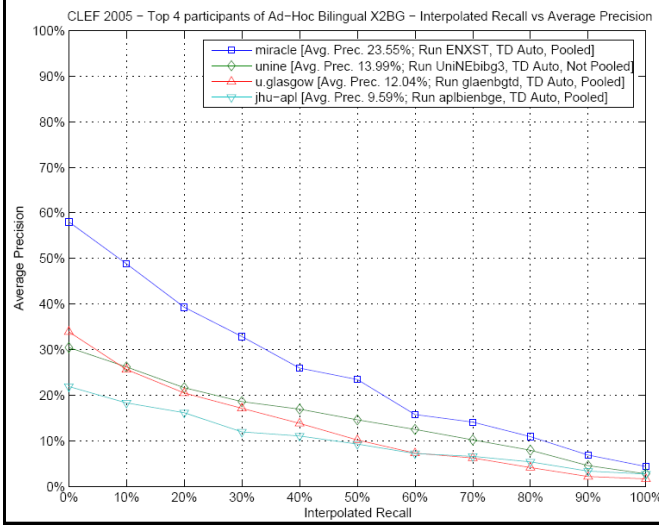



- Known language
- Various translation tools available

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

177

# 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

178

# 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
  - MT system available
  - Parallel or comparable corpora

179

## Outline

- Information Retrieval
- MLIA/CLIR motivation and evaluation campaigns
- Indexing
- Translation
- **Matching**
- Demo
- Lab Exercise

180

## 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.)

## Word Statistics

- Weighting schemes addressing these assumptions need word statistics:
  - *ff* (feature frequency), for text equiv. *tf* (term frequency) – number of occurrences of a feature/term in a document
  - *df* (document frequency) – number of documents with a feature/term
  - document length – measure for the length of a document: number of tokes, number of features, byte length
  - positional information

## Inverse Document Frequency

- The "*Inverse Document Frequency* idf($\varphi_k$)" captures the "rareness" of a term:
- **idf($\varphi_k$) = log((1+N)/(1+df($\varphi_k$)))**
- where
  - N: number of documents in a collection
  - df($\varphi_k$): number of documents that contain term k

## Inverse Document Frequency

- Which terms are "characteristic" of a document?
  - **$w(\varphi_k,d_i) = tf(\varphi_k,d_i) \bullet idf(\varphi_k)$**
  - where
    - $tf(\varphi_k,d_i)$: number of occurrences of term k in document $d_i$
    - $idf(\varphi_k)$ : idf of term k
  - Terms with a high weight according to this formula are frequent in a specific document, but rare in the overall collection.

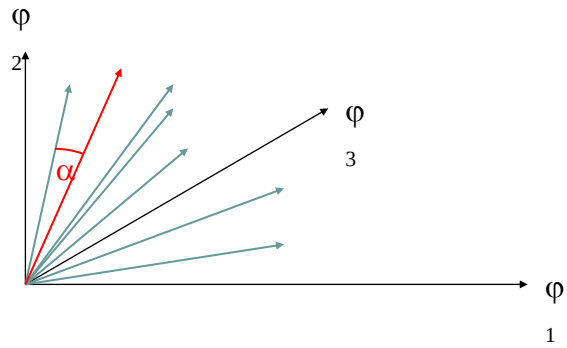
## Vector Space Model

- We now have an idea on how to weigh individual terms. But we have to weigh the query as a whole.
- Retrievable items (documents) $d_j$ and the query q are vectors in an high-dimensional feature space.

  Vector dj =(...,$w_{\varphi i}$,...) (Weight $w_{\varphi i}$, e.g. binary, or according to tf.idf)

# Vector Space Model

- Documents are vectors
- Query are vectors
- Vector space has dimensionality of n = number of different features in the collection
- Sim(Document, Query) = Cosine(Angle)




# tf·idf - Cosine

- We obtain the following weigthing formula when using tf.idf-weights for individual features

$$a_{i,j} := ff(\varphi_i, d_j) * idf(\varphi_i)$$

$$b_i := ff(\varphi_i, q) * idf(\varphi_i)$$

$$RSV(q, d_j) := \frac{\sum_{\varphi_i \in \Phi(q) \cap \Phi(d_j)} a_{i,j} * b_i}{\sqrt{\sum_{\varphi_i \in \Phi(d_j)} a_{i,j}^2} * \sqrt{\sum_{\varphi_i \in \Phi(q)} b_i^2}}$$

- This is a well-known, "classical" formula
- But we can do better...

# Okapi

- The document score is computed with parameter (*b* (=0.75), $k_1$ (=1.2)) as:

$$RSV(D_i, Q) = \sum_{j=1}^{t} \frac{(k_1+1)\cdot tf_{ij}}{K+tf_{ij}} \cdot w^{(1)} \cdot tf_{qj}$$

$$\text{with } K = k_1 \cdot \left[(1-b) + \frac{b \cdot l(D_i)}{avdl}\right] \qquad w^{(1)} = \log\left(\frac{n - df_j + 0.5}{df_j + 0.5}\right)$$

- This probabilistic model returns one of the best MAP (on the Web, TREC, or multilingual corpora)

The parameters b and $k_1$ could be adjusted depending on the collection

188

# Divergence from Randomness

- Advanced parametric probabilistic model (Amati & van Rijsbergen (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness », *ACM TOIS*, 20(4), p. 357-389)
- Combining two aspects
  1. $Prob^1$ is the pure chance probability of finding *tf* occurrences of the indexing unit in the document (informative content). If $Prob^1$ is high (the term is randomly distributed, it brings little information)
  2. $Prob^2$ is the probability of encountering a new occurrence of the given term in the document given that we have already found *tf* occurrences of this indexing unit. (1-$Prob^2$ is the first normalization of the information content)

$$\begin{aligned} w_{ij} &= Inf^1_{ij}(tf) \cdot Inf^2_{ij}(tf) \\ &= -\log_2[Prob^1_{ij}(tf)] \cdot [1 - Prob^2_{ij}(tf)] \end{aligned}$$

189

# Divergence from Randomness

- **P**oisson (approx. Bernoulli model) with λ the mean number of occurrences per document

$$Prob_{ij}^{1}(tf) = \frac{e^{\lambda_j} \cdot \lambda_j^{tf_{ij}}}{tf_{ij}!} \quad \text{with } \lambda_j = \frac{tc_j}{n}$$

- **G**eometric distribution (with p=1/(1+λ))

$$Prob_{ij}^{1}(tf) = \left(\frac{1}{1+\lambda_j}\right) \cdot \left(\frac{\lambda_j}{1+\lambda_j}\right)^{tf} \quad \text{with } \lambda_j = \frac{tc_j}{n}$$

190

# Divergence from Randomness

- **I**DF model

  Compute the probability of choosing a document having one occurrence of the underlying term. Having *n* documents in the corpus, and *df* documents having (at least) one occurrence of the corresponding term,

$$Prob_{ij}^{1}(tf = 1) = \left(\frac{df_j + 0.5}{n+1}\right)$$

$$Prob_{ij}^{1}(tf) = \left(\frac{df_j + 0.5}{n+1}\right)^{tf}$$

191

# Divergence from Randomness

- Prob$^2$ (for the elite set, the set of documents in which thee underlying term occur) could be estimated using Laplace law

$$Prob[tf_{ij}+1|tf_{ij}, D_i] \;=\; Prob^2_{ij}(tf) \;=\; \frac{tf_{ij}+1}{tf_{ij}+2} \;\approx\; \frac{tf_{ij}}{tf_{ij}+1}$$

- **B**ernoulli (ratio of two Bernoulli processes)

$$Prob^2_{ij}(tf) \;=\; \frac{tc_j+1}{n\cdot(tf_{ij}+1)}$$

192

# Divergence from Randomness

- To be effective, the *tf* component must be normalized in order to take account for the size difference of documents

$$tfn_{ij} = tf_{ij}\cdot\left(\frac{avdl}{l(d_i)}\right) \quad \text{or}$$

$$tfn_{ij} = tf_{ij}\cdot\log_2\left(1+\frac{avdl}{l(d_i)}\right)$$

replace *tf* by *tfn* in the previous equations (the second is more effective).

193

## Matching (Overview)

- Tokenization (only, T queries, e.g. "Pesticides")
  - MAP: 0.3611 (based on 284 queries)
- + remove stopword list (e.g. "pesticides")
  - MAP: 0.3743 (+3.7%;  improve: 149, hurt: 97, same: 38)
- + SMART (e.g. "pesticid")
  - MAP: 0.4152 (13.2%;  improve: 153, hurt: 98, same: 33)
- or + S-stemmer (e.g. "pesticide")
  - MAP: 0.4044 (+12.0%;  improve: 157, hurt: 99, same: 28)
- S-stemmer & pseudo-relevance Feedback (Rocchio) 5 / 20 (e.g. "pesticide public cancer environmental")
  - MAP: 0.4314 (+19.5%;  improve: 166, hurt: 93, same: 25)

194

## Matching (Stem vs. *n*-gram)

- Tokenization (only, T queries, e.g., "pesticides")
  - MAP: 0.3611 (based on 284 queries)
- + remove stopword list (e.g., "pesticides")
  - MAP: 0.3743 (+3.7%; improve: 149, hurt: 97, same: 38)
- + S-stemmer (e.g., "pesticide")
  - MAP: 0.4044 (12.0%;  improve: 157, hurt: 99, same: 28)
- 5-gram (e.g., "pesti estic stici ticid icide cides")
  - MAP: 0.3622 (+0.3%;  improve: 94, hurt: 167, same: 23)
- trunc(5) (e.g., "pesti")
  - MAP: 0.4081 (+13.0%;  improve: 128, hurt: 134, same: 22)

195

## Matching (Query Translation)

- English T queries (S-stemmer, Okapi, e.g., "Pesticides")
  - MAP: 0.4044
- German T queries (Google MT, e.g., "Pestizide")
  - MAP: 0.3625 (-10.4%; improve: 42, hurt: 88, same: 154)
- Spanish T queries (Google MT, e.g., "Pesticidas")
  - MAP: 0.3752 (-7.2%; improve: 40, hurt: 78, same: 166)
- French T queries (Google MT, e.g., "Des pesticides")
  - MAP: 0.3692 (-8.7%; improve: 56, hurt: 84, same: 144)
- Chinese T queries (Google MT, "嬰兒食品中含有殺蟲劑")
  - MAP: 0.3327 (-17.7%; improve: 68, hurt: 117, same: 99)

196

## 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

197

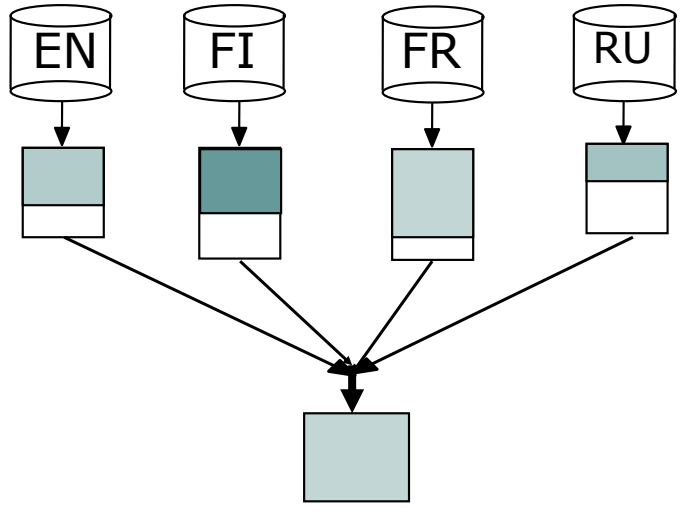
# Multilingual IR

- 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)

198

# Multilingual IR (QT)




199

# Multilingual IR

Merging problem

| | | |
|---|---|---|
| 1 EN120 1.2 | 1 FR043 0.8 | 1 RU050 6.6 |
| 2 EN200 1.0 | 2 FR120 0.75 | 2 RU005 6.1 |
| 3 EN050 0.7 | 3 FR055 0.65 | 3 RU120 3.9 |
| 4 EN705 0.6 | 4 … | 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]

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

202

# 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** |

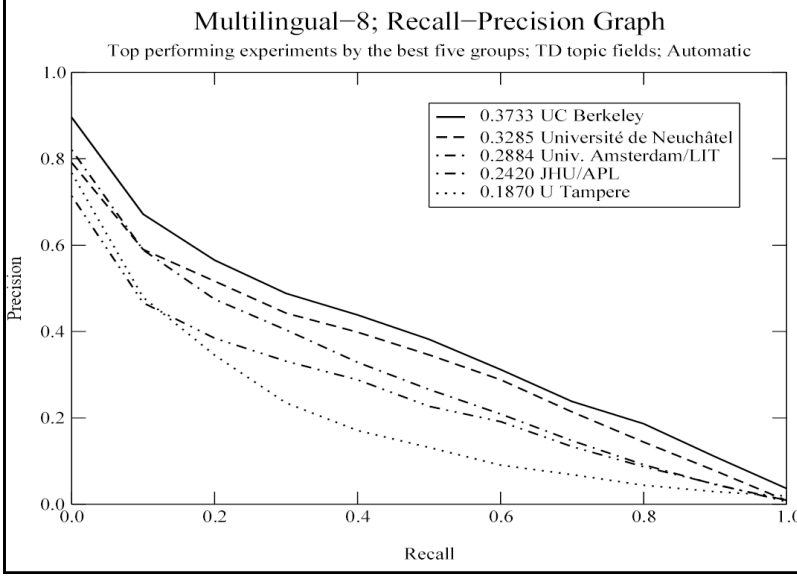
203

# 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)

204

# Multilingual IR (CLEF-2003)




205

# 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)

206

# 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)

207

## 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*)

208

## The Future

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

209

# General References

- Conference
  - ACM-SIGIR
  - ECIR
- 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
- Manning, C.D., Raghavan, P., Schütze, H. *Introduction to Information Retrieval.* Cambridge University Press, 2008.
- Croft, W.B., Metzler, D., Strohman, T. *Search Engines Information Retrieval in Practice*. Pearson, 2010

210

# References


Abdou, S., Savoy, J. 2006. Statistical and comparative evaluation of various indexing and search models. In *AIRS-2006*, October 2006, Singapore, Springer-Verlag, Berlin, LNCS #4182, 362-373.

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.

Ballesteros, L., Croft, B.W. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. *ACM SIGIR'97*, 84-91.

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. *IR Journal*, 7(1-2), 7-31.

Braschler, M. 2004. Combination approaches for multilingual text retrieval. *IR Journal*, 7(1-2), 183-204.


211

# References


Dolamic, L., 2009. Influence of the Language Morphology on Information Retrieval. PhD thesis.

Fox, C. 1990. A stop list for general text. *ACM-SIGIR Forum*, 24(1):19-35.

Gale, W.A., Church, K.WGao, 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.

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.


212

# References


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

Kettunen, K. 2009. Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval. *Journal of Documentation, 65(2),* 267-290.

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.

McNamee, P. 2008. Textual Representations for Corpus-Based Bilingual Retrieval. PhD Thesis, John Hopkins University.


213

# References


McNamee, P., Nicholas, C., Mayfield, J. 2009. Addressing Morphological Variation in Alphabetic Languages. *ACM-SIGIR 2009*.

MultiMatch project (www.multimatch.org)

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.

Peat, H. J., Willett, P. 1991. The limitations of term co-occurrence data for query expansion in document retrieval systems. *JASIS, 42(5), 1991, 378-383*

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.




# References


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

Savoy J. 2008. Searching Strategies for the Bulgarian Language. *IR Journal*, 10(6), 2008, 509-529.

Savoy J. 2008. Searching Strategies for the Hungarian Language. *Information Processing & Management*, 44(1), 2008, 310-324.

Savoy J., Dolamic, L. 2009. How effective is Google's translation service in search?. *Communications of the ACM*, 2009, to appear.

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., Nguyen, 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

# Web Resources

Official Journal of EU: eur-lex.europa.eu

United Nations: www.un.org

EuroWordNet: www.illc.uva.nl/EuroWordNet/

DGT (UE): ec.europa.eu/translation/

Evaluation campaigns: CLEF, NTCIR, TREC

http://romip.ru/en (in Russian language only)

Trésor de la langue française: atilf.atilf.fr/tlf.htm