Multilingual Information Retrieval

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

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

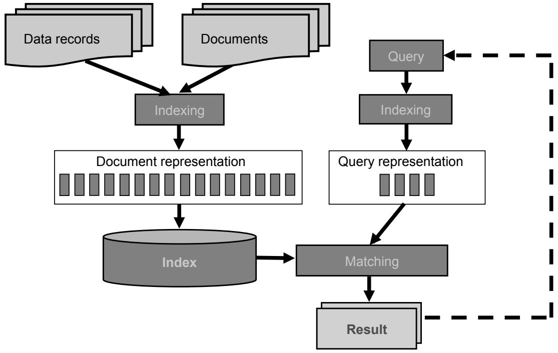
Outline

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

Information Retrieval (IR)

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

IR Flow



The MLIR Challenge

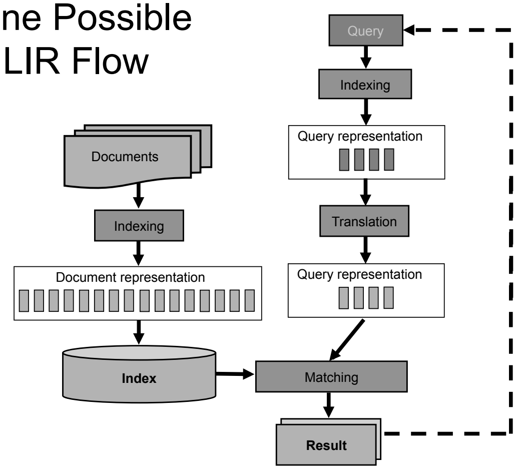
"Given a query in *any medium* and *any language*, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or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

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

MLIR / CLIR

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

One Possible MLIR Flow



Query

Indexing

Query representation

Translation

Query representation

Documents

Indexing

Document representation

Index

Matching

Result

MLIR Reality

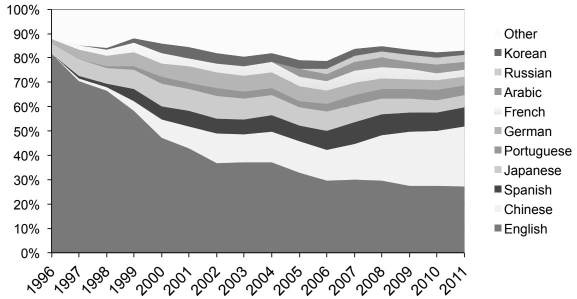
- Strč prst skrz krk
- Mitä sinä teet?
- Mam swoją książkę
- Nem fáj a fogad?
- Er du ikke en riktig nordmann?
- Добре дошли в България!
- Fortuna caeca est
- 我不是中国人

MLIR Reality

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

MLIR Reality

Internet users by language (www.internetworldstat.com)



MLIR Reality

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

Evaluation Campaigns

- TREC (trec.nist.gov)
  - 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)

Evaluation Campaigns

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

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

Evaluation Campaigns

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

Evaluation Methodology

- Compare retrieval performance using a test collection
- To compare *relatively* the performance of two techniques:
  - each technique used to evaluate test queries
  - results (set or ranked list) compared using some performance measure
  - most common measures - *precision* and *recall*
- Effectiveness measure
  - MAP Mean Average Precision
  - MRR Mean Reciprocal Rank
- Statistical testing is required

Outline

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

Indexing

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

Indexing Step 1: Preprocessing

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

What does that means for non-English IR?

Beyond Just English

<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

- Alphabets
  - Latin alphabet (26)
  - Cyrillic (33): спутник
  - Arabic (28), Hebrew
  - Other Asian languages: Hindi, Thai
- Syllabaries
  - Japan: Hiragana (46) における
    Katakana (46) 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

Beyond Just English

- Encoding systems
  - ASCII is limited to 7 bits
  - Windows, Macintosh, BIG5, GB, EUC-JP, EUC-KR, …
  - ISO-Latin-1 (ISO 8859-1 West European), Latin-2 (East European), Latin-3 (South European), Latin-4 (North European), Cyrillic (ISO-8859-5), Arabic (ISO-8859-6),…
  - Unicode (UTF-8, see www.unicode.org)
- One language ≠ one encoding
- Input / output devices (at least the query)
- Tools
  - What is the result of a `sort` on Japanese words?

Even English is not Just English

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

Indexing Step 2: Identification

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

Language Identification

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

Indexing Step 3: Granularity

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

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

Indexing Step 4: Segmentation

- The document is split into "valid" tokens
  "To be or not to be" 6 tokens, but 4 word types

- The tokens are suitable to form the index structure
- "Undesirable" tokens are eliminated
  - non-content bearing tokens
  - special characters
  - numbers, date, amounts in $
  - very short or very long tokens, ...

Segmentation

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

Segmentation

- Compound construction
  Morphological characteristic used by many languages
  - EN: handgun, viewfinder
  - FR: “porte-clefs” (key ring) "chemin de fer" (railway)
  - IT: “capoufficio” (chief of the office) = "capo" + "ufficio"
    but "capiufficio" (plural)
    but "capogiro" (sing) and "capogiri" (plural) (dizinesss)
  - BU: “радиоапарат” = “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

Segmentation

- For the German language
  - In DE: “Bundesbankpräsident” =
    “Bund” + es + “Bank” + “Präsident”
    federal bank CEO
  - Different forms in the queries and documents
    “Computersicherheit”
    could appear as
    “die Sicherheit mit Computern”
  - Automatic decompounding is useful (+23% in MAP, short queries, +11% longer queries, [Braschler & Ripplinger 2004].

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))

32

Segmentation

- Language independent approach
  n-gram indexing [McNamee & Mayfield 2004], [McNamee 2008]
  - 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
  - trunc-*n*, consider only the first *n* letters
    compute → “compu“

33

Segmentation

A Chinese sentence, various representations

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

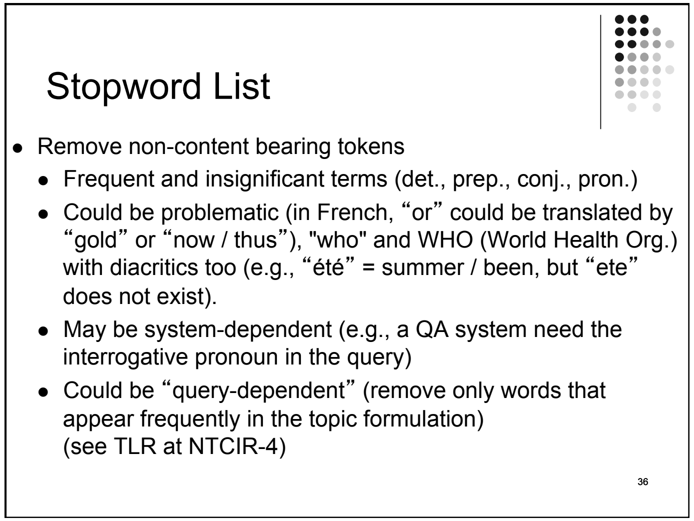
34

Segmentation in ZH

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

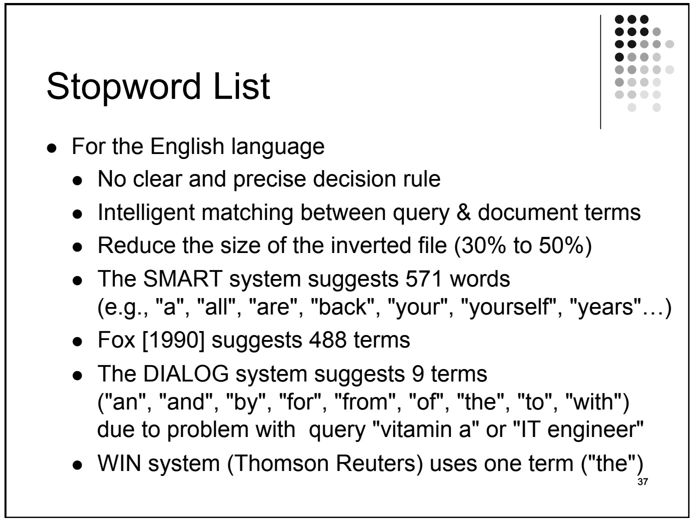
MAP / ZH (T) NTCIR-5	unigram	bigram	word (MTool)	uni+ bigram
PB2	0.2774	**0.3042**	0.3246	<u>0.3433</u>
LM	0.2995	**0.2594**	0.2800	0.2943
Okapi	0.2879	**0.2995**	0.3231	<u>0.3321</u>
tf idf	<u>0.1162</u>	**0.2130**	<u>0.1645</u>	0.2201

35



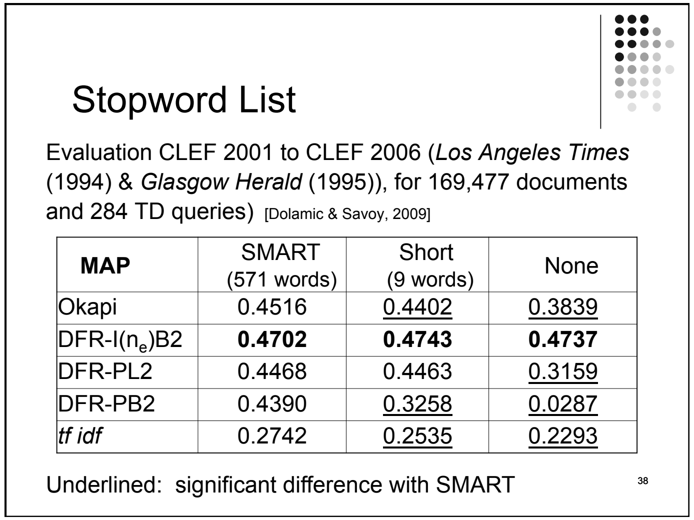
Stopword List

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



Stopword List

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

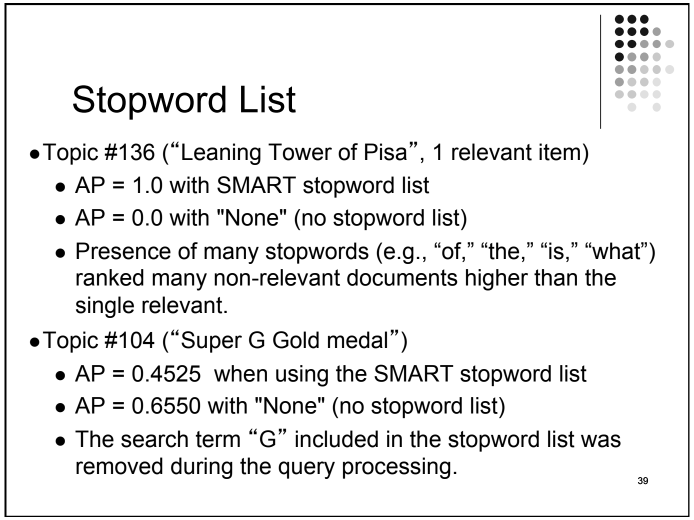


Stopword List

Evaluation CLEF 2001 to CLEF 2006 (*Los Angeles Times* (1994) & *Glasgow Herald* (1995)), for 169,477 documents and 284 TD queries) [Dolamic & Savoy, 2009]

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



Stopword List

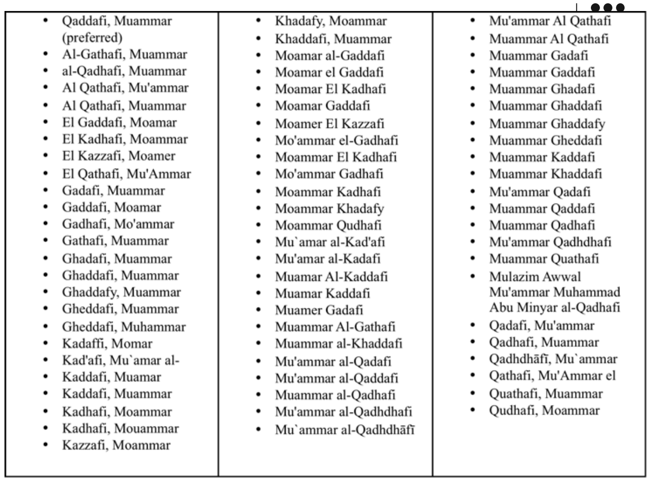
- Topic #136 (“Leaning Tower of Pisa”, 1 relevant item)
  - AP = 1.0 with SMART stopword list
  - AP = 0.0 with "None" (no stopword list)
  - Presence of many stopwords (e.g., “of,” “the,” “is,” “what”) ranked many non-relevant documents higher than the single relevant.
- Topic #104 (“Super G Gold medal”)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

Indexing Step 5: Normalization

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

Normalization

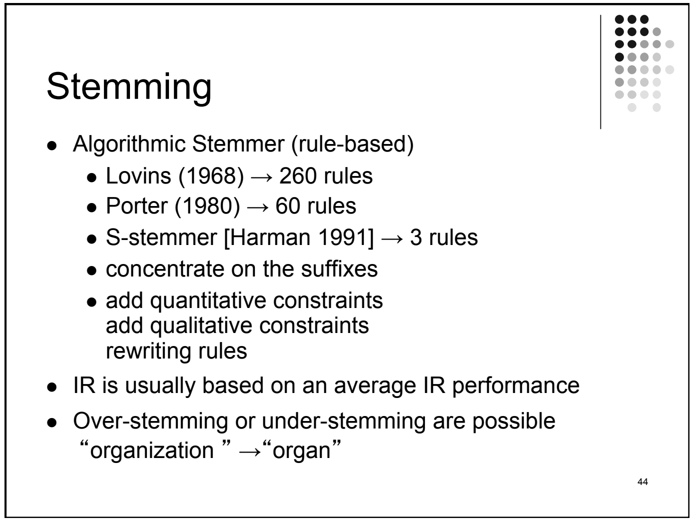
- Diacritics
  - differ from one language to another (“résumé”, “Äpfel”)
  - could be used to distinguish the meaning (e.g., “tache” (task) or “tâche (mark, spot))
- Normalization / Proper nouns
  - Spelling may change with languages
    Gorbachev, Gorbacheff, Gorbachov
    Mona Lisa ↔ La Joconde ↔ La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



- Qaddafi, Muammar (preferred)
- Al-Gathafi, Muammar
- al-Qadhafi, Muammar
- Al Qathafi, Mu'ammar
- Al Qathafi, Muammar
- El Gaddafi, Moamar
- El Kadhafi, Moammar
- El Kazzafi, Moamer
- El Qathafi, Mu'Ammar
- Gadafi, Muammar
- Gaddafi, Moamar
- Gadhafi, Mo'ammar
- Gathafi, Muammar
- Ghadafi, Muammar
- Ghaddafi, Muammar
- Ghaddafy, Muammar
- Gheddafi, Muammar
- Gheddafi, Muhammar
- Kadaffi, Momar
- Kad'afi, Mu`amar al-
- Kaddafi, Muamar
- Kaddafi, Muammar
- Kadhafi, Moammar
- Kadhafi, Mouammar
- Kazzafi, Moammar
- Khadafy, Moammar
- Khaddafi, Muammar
- Moamar al-Gaddafi
- Moamar el Gaddafi
- Moamar El Kadhafi
- Moamar Gaddafi
- Moamer El Kazzafi
- Mo'ammar el-Gadhafi
- Moammar El Kadhafi
- Mo'ammar Gadhafi
- Moammar Kadhafi
- Moammar Khadafy
- Moammar Qudhafi
- Mu`amar al-Kad'afi
- Mu'amar al-Kadafi
- Muamar Al-Kaddafi
- Muamar Kaddafi
- Muamer Gadafi
- Muammar Al-Gathafi
- Muammar al-Khaddafi
- Mu'ammar al-Qadafi
- Mu'ammar al-Qaddafi
- Muammar al-Qadhafi
- Mu'ammar al-Qadhdhafi
- Mu`ammar al-Qadhdhāfī
- Mu'ammar Al Qathafi
- Muammar Al Qathafi
- Muammar Gadafi
- Muammar Gaddafi
- Muammar Ghadafi
- Muammar Ghaddafi
- Muammar Ghaddafy
- Muammar Gheddafi
- Muammar Kaddafi
- Muammar Khaddafi
- Mu'ammar Qadafi
- Muammar Qaddafi
- Muammar Qadhafi
- Mu'ammar Qadhdhafi
- Muammar Quathafi
- Mulazim Awwal Mu'ammar Muhammad Abu Minyar al-Qadhafi
- Qadafi, Mu'ammar
- Qadhafi, Muammar
- Qadhdhāfī, Mu`ammar
- Qathafi, Mu'Ammar el
- Quathafi, Muammar
- Qudhafi, Moammar

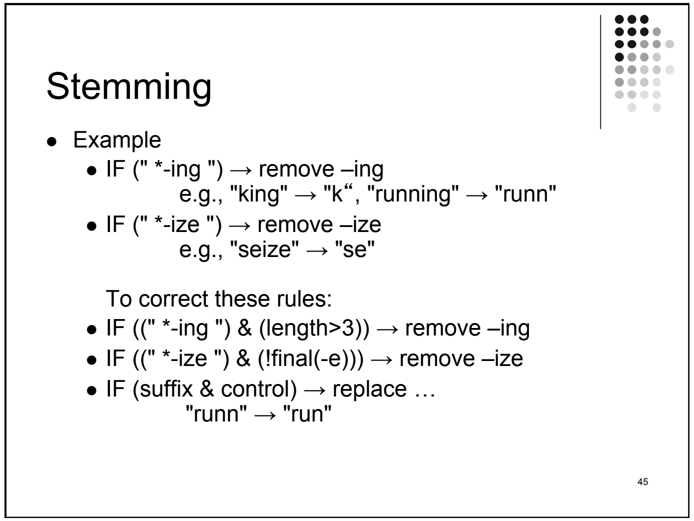
Normalization

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



Stemming

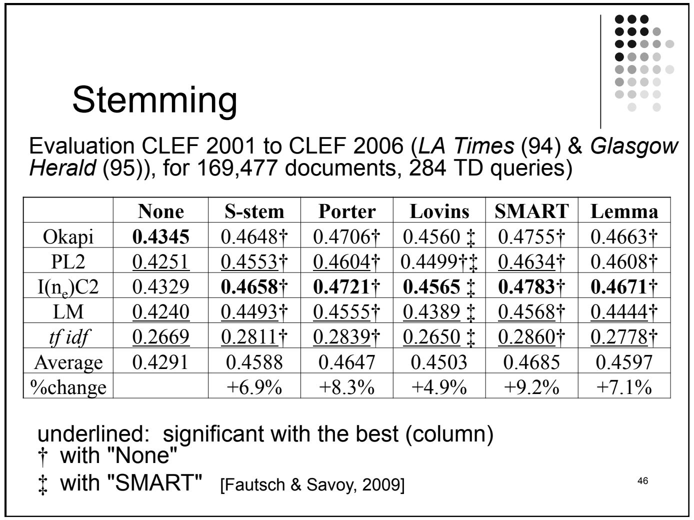
- Algorithmic Stemmer (rule-based)
  - Lovins (1968) → 260 rules
  - Porter (1980) → 60 rules
  - S-stemmer [Harman 1991] → 3 rules
  - concentrate on the suffixes
  - add quantitative constraints
    add qualitative constraints
    rewriting rules
- IR is usually based on an average IR performance
-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



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"

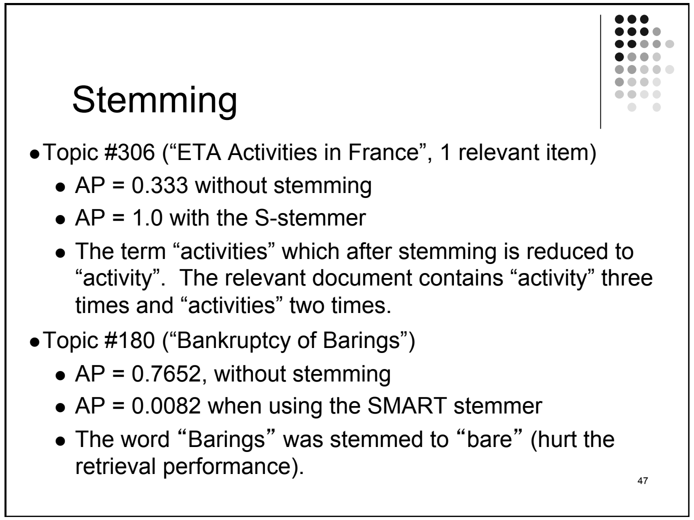


Stemming

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

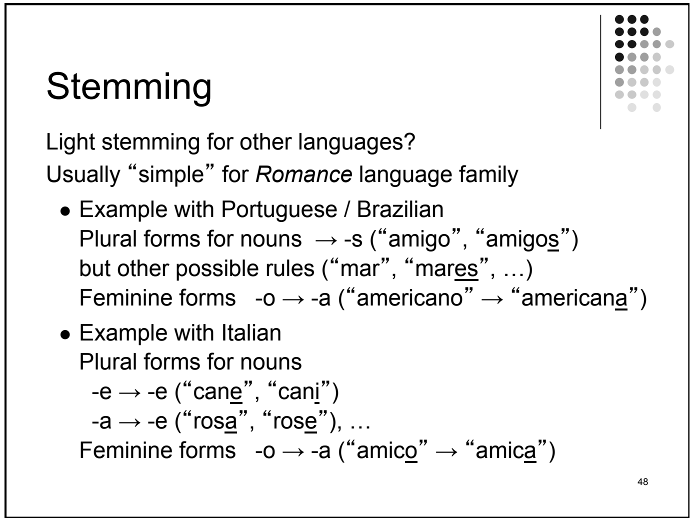
	None	**S-stem**	**Porter**	**Lovins**	**SMART**	**Lemma**
Okapi	**0.4345**	0.4648†	0.4706†	0.4560 ‡	0.4755†	0.4663†
PL2	0.4251	0.4553†	0.4604†	0.4499†‡	0.4634†	0.4608†
I(n_e)C2	0.4329	**0.4658†**	**0.4721†**	**0.4565 ‡**	**0.4783†**	**0.4671†**
LM	0.4240	0.4493†	0.4555†	0.4389 ‡	0.4568†	0.4444†
tf idf	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" [Fautsch & Savoy, 2009]



Stemming

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

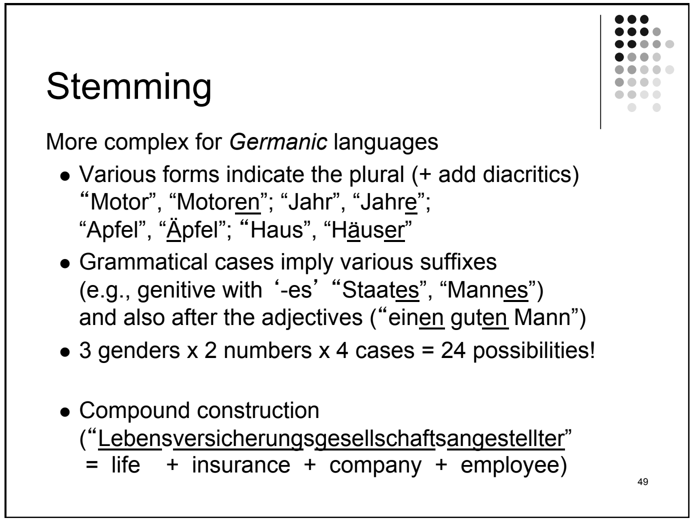


Stemming

Light stemming for other languages?

Usually "simple" for *Romance* language family

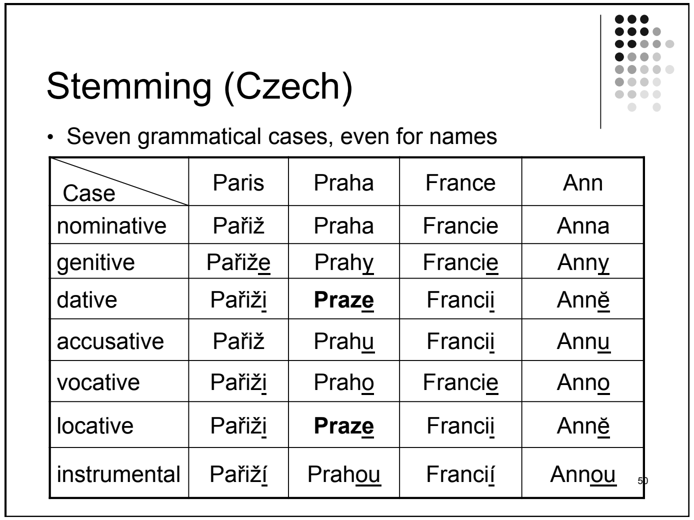
- Example with Portuguese / Brazilian
  Plural forms for nouns → -s ("amigo", "amigos")
  but other possible rules ("mar", "mares", …)
  Feminine forms -o → -a ("americano" → "americana")
- Example with Italian
  Plural forms for nouns
  -e → -e ("cane", "cani")
  -a → -e ("rosa", "rose"), …
  Feminine forms -o → -a ("amico" → "amica")



Stemming

More complex for *Germanic* languages

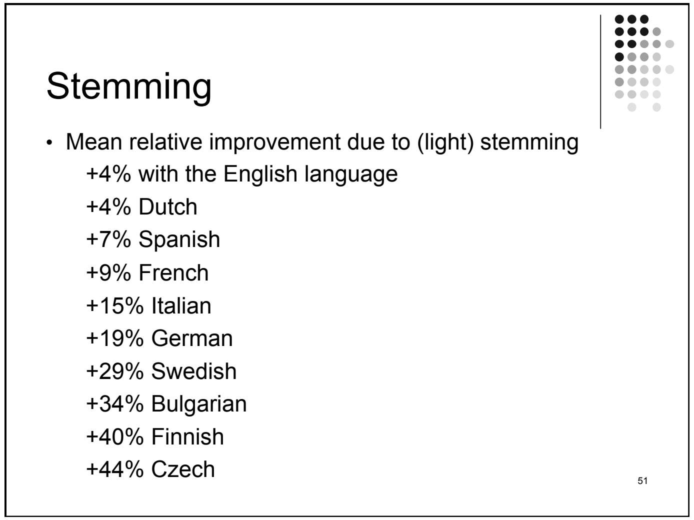
- Various forms indicate the plural (+ add diacritics)
  "Motor", "Motoren"; "Jahr", "Jahre";
  "Apfel", "Äpfel"; "Haus", "Häuser"
- Grammatical cases imply various suffixes
  (e.g., genitive with '-es' "Staates", "Mannes")
  and also after the adjectives ("einen guten Mann")
- 3 genders x 2 numbers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



Stemming (Czech)

- Seven grammatical cases, even for names

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



Stemming

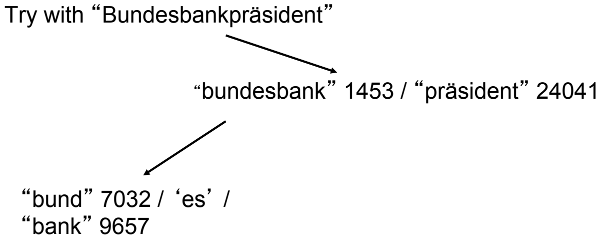
- Mean relative improvement due to (light) stemming
  +4% with the English language
  +4% Dutch
  +7% Spanish
  +9% French
  +15% Italian
  +19% German
  +29% Swedish
  +34% Bulgarian
  +40% Finnish
  +44% Czech

Decompounding (German)

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

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**		

Decompounding (German)



Try with “Bundesbankpräsident”

“bundesbank” 1453 / “präsident” 24041

“bund” 7032 / ‘es’ /
“bank” 9657

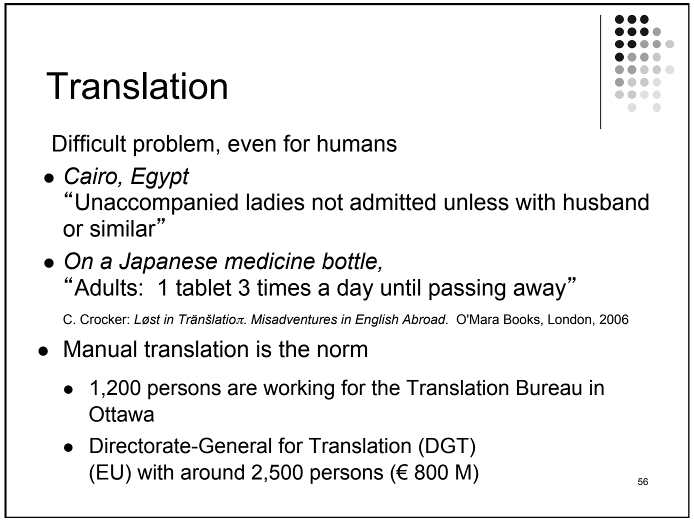
A similar issue with compounds also exists in other Germanic languages, such as Dutch, Swedish, ... as well as other languages (Hungarian)

Indexing Step 6: Enrichment

- Documents are enriched with extra features, or with more specialised features
  - Named Entity recognition
  - Thesauri for expansion
  - Anchor text from inlinks
  - Contextual information (from user profiles, from linked pages, from clustering, ...)
  - ...

Outline

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



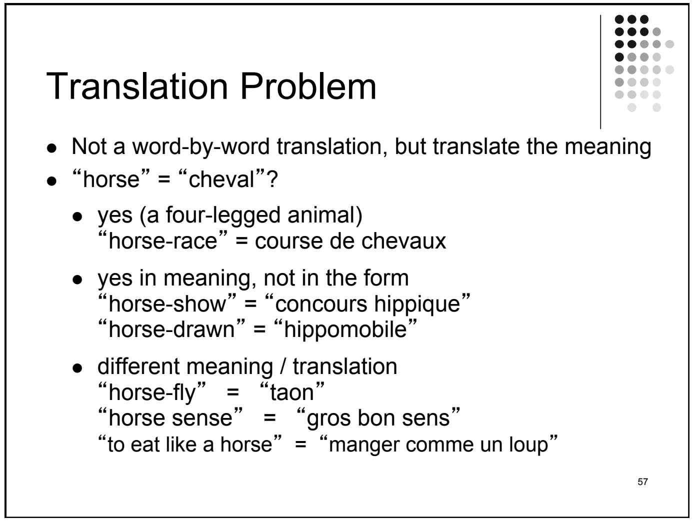
Translation

Difficult problem, even for humans

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

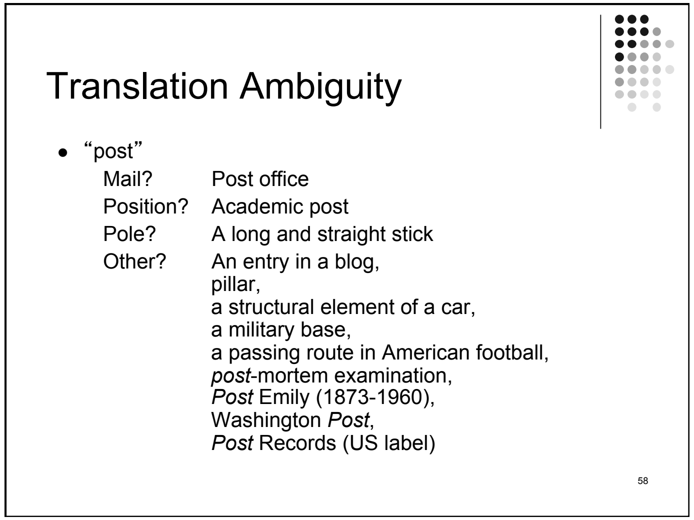
C. Crocker: *Løst in Tränšlatioπ. Misadventures in English Abroad.* O'Mara Books, London, 2006

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



Translation Problem

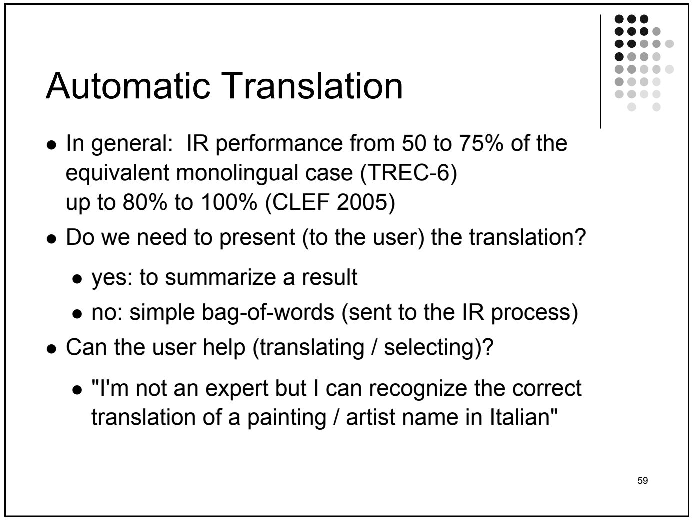
- Not a word-by-word translation, but translate the meaning
- “horse” = “cheval”?
  - yes (a four-legged animal)
    “horse-race” = course de chevaux
  - yes in meaning, not in the form
    “horse-show” = “concours hippique”
    “horse-drawn” = “hippomobile”
  - different meaning / translation
    “horse-fly”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



Translation Ambiguity

- “post”

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



Automatic Translation

- In general: IR performance from 50 to 75% of the equivalent monolingual case (TREC-6)
  up to 80% to 100% (CLEF 2005)
- Do we need to present (to the user) the translation?
  - yes: to summarize a result
  - no: simple bag-of-words (sent to the IR process)
- Can the user help (translating / selecting)?
  - "I'm not an expert but I can recognize the correct translation of a painting / artist name in Italian"

Automatic Translation

- In many cases, the context could be rather short
  - Query translation
    could be a mix of bag-of-words and phrase
    E.g., “orange plate with a table”
    difficult to understand/classify
    “orange plate” a noun phrase or a bag of words
  - Legend of statistical tables
  - Caption of images
  - Short description of a cultural object
    (with a mixed of languages, e.g., *The European Library*)

Translation Strategies

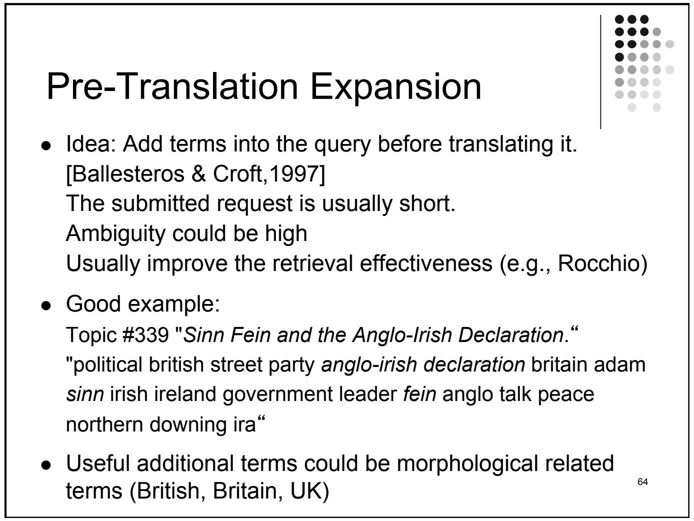
- Ignore the translation problem!
  Sentence in one language is misspelled expression of the other (near cognates) and with some simple matching rules, a full translation is not required
  (e.g., Cornell at TREC-6, Berkeley at NTCIR-5)
- Machine-readable bilingual dictionaries (MRD)
  - provide usually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

OOV

- Out-Of-Vocabulary
  - Dictionary has a limited coverage (both in direct dictionary-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?

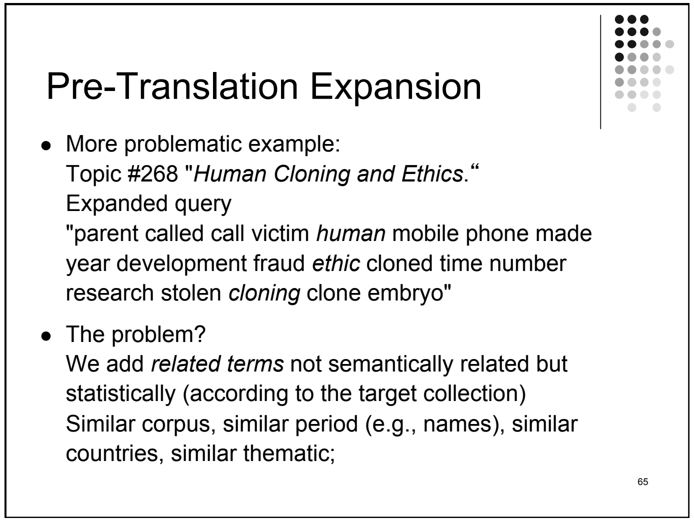
Translation Strategies

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



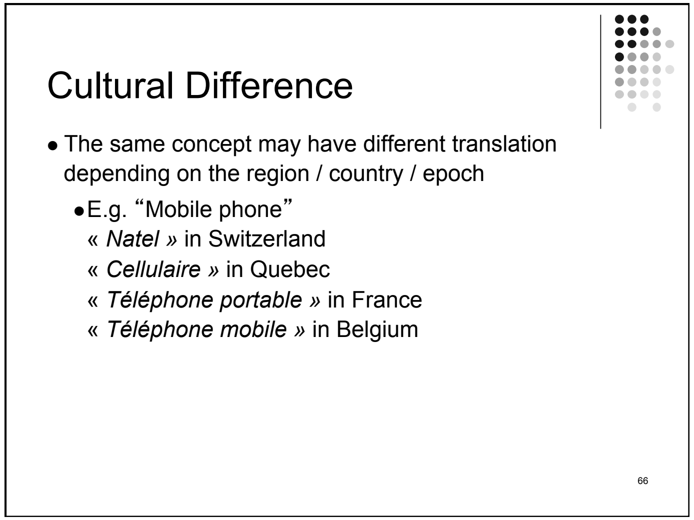
Pre-Translation Expansion

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



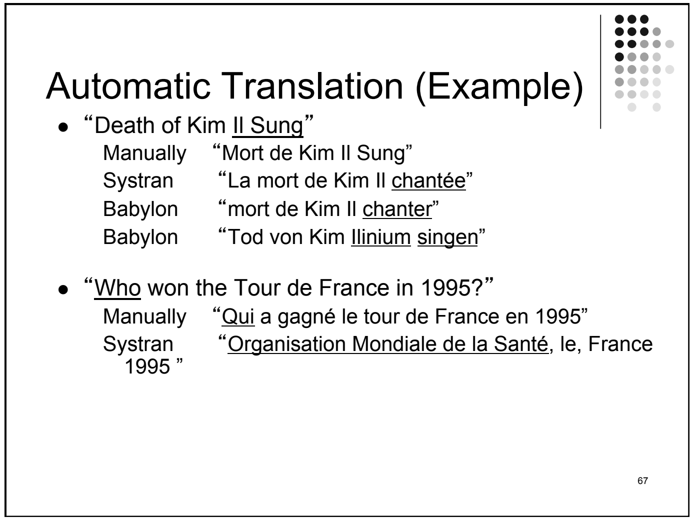
Pre-Translation Expansion

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



Cultural Difference

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



Automatic Translation (Example)

- “Death of Kim Il Sung”

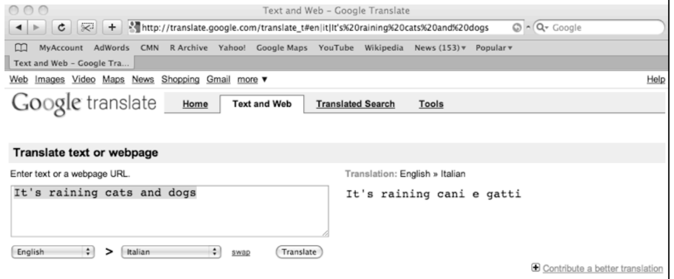
Manually	“Mort de Kim Il Sung”
Systran	“La mort de Kim Il chantée”
Babylon	“mort de Kim Il chanter”
Babylon	“Tod von Kim Ilinium singen”

- “Who won the Tour de France in 1995?”

Manually	“Qui a gagné le tour de France en 1995”
Systran	“Organisation Mondiale de la Santé, le, France 1995 ”

Automatic Translation (Example)

- Example EN → IT (idiomatic)



Translate text or webpage

Enter text or a webpage URL.

It's raining cats and dogs

Translation: English » Italian

It's raining cani e gatti

English > Italian swap Translate

Contribute a better translation

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

Translation

On a large query set (284 CLEF 2001-06, English corpus)

Original query written in English (Title-only) [Dolamic & Savoy 2010b]

Statistical significant difference (*)

	MAP
	Mono
I(ne)C2	**0.4053**
Okapi	0.4044
LM	0.3708*
tf idf	0.2392*

Translation

Original query written in English (284 T-only)

Automatic translation done by Google (May 2007)

Statistical significant difference (*) [Dolamic & Savoy 2010b]

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%

Translation

Original query written in English (284 T-only)

Automatic translation done by Yahoo (may 2007)

Statistical significant difference (*) [Dolamic & Savoy 2010b]

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%

Translation Strategies

Some findings

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

Translation

Where are the real translation problems?

For Google MT system

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

Outline

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

Matching: Assumptions

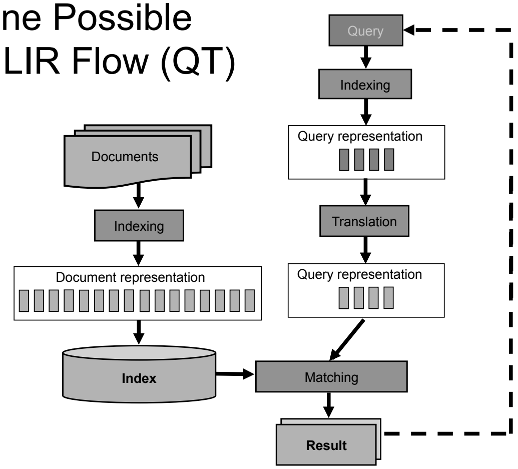
- The matching stage needs to assign weights to query (and document) terms
- Remember: we should not require exact matches
- Assumptions:
  - Texts having similar vocabulary tend to have the same meaning
  - More query terms match → more relevant
  - Query terms more frequent in doc → more relevant
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

Multililingual IR

- If I need to add one language?

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

One Possible MLIR Flow (QT)



MLIR - Query Translation

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

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

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

MLIR - Query Translation

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

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

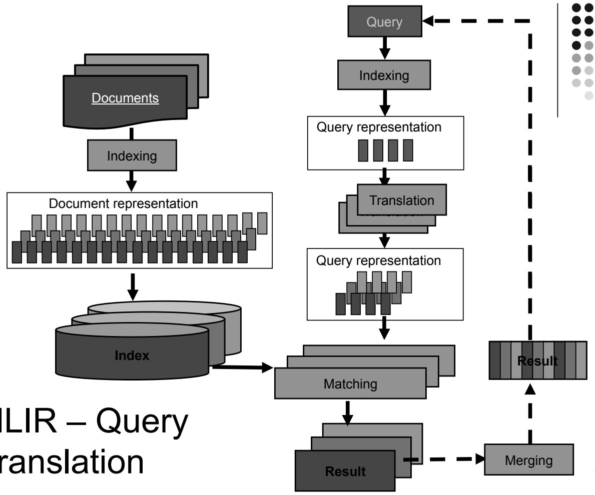
Moreover, the corpus C (in the source language) could be different (thematic, time, geographic, etc.) than the corpus in the target language (used by the D and denoted C_l). We may estimate as:

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

MLIR – Query Translation

- If I need to consider *more* than one language?

- More complex setup
- A series of bilingual steps
  Query translation (QT) and search into the different languages, then merging
  - Translate the query into different languages
  - Perform a search separately into each language
  - Merge the result lists



MLIR – Query Translation

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

- 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]

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

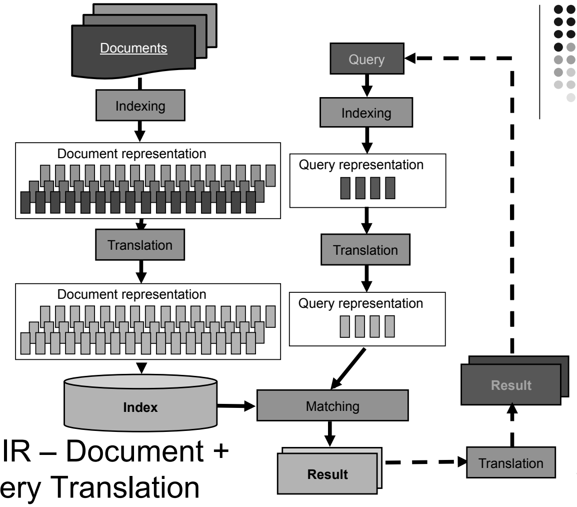
Multilingual IR

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**

MLIR – Document Translation

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



MLIR – Document + Query Translation

Multilingual IR

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

Multilingual IR

- Using QT approach and merging: simplicity
  - Logistic regression work well
  - Normalization is usually good (e.g., Z-score or by max)
  - But when using the same IR system, raw-score merging (simple) could offer an high level of performance
  - For better merging method see CMU at CLEF 2005
- Using DT: 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)

Conclusion

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

Conclusion

- "*In theory, practice and theory are the same, but in practice they are not*.“
  David Hawking, Chief Scientist *Funnelback*
- The various experiments shown that query-by-query analysis is an important step in scientific investigations. We really need to understand why IR system may (will) fail for some topics. Learn by experiences.
- The real problems (implementation) are crucial
  (*Der Teufel liegt im Detail*)
- ”*Words come and go. Grammar fluctuates. Pronunciations alter. Spelling preferences vary*.“
  David Crystal

References

- Conference
  - ACM-SIGIR
  - ECIR
  - AIRS
- Journal
  - Information Retrieval Journal, IRJ (Springer)
  - Information Processing & Management, IP&M (Elsevier)
  - Journal of the American Society for Information Science & Technology, JASIST (Wiley)
- Evaluation campaigns: CLEF, NTCIR, TREC, FIRE

References

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.

Dolamic L., Savoy J. (2010). When Stopword Lists Make the Difference. *Journal of the American Society for Information Sciences and Technology*, 61(1), 200-203

Dolamic L., Savoy J. (2010b). Retrieval Effectiveness of Machine Translated Queries. *Journal of the American Society for Information Sciences and Technology*, to appear

References

Fautsch C., Savoy J. (2009). Algorithmic Stemmers or Morphological Analysis: An Evaluation. *Journal of the American Society for Information Sciences and Technology*, 60(8), 1616-1624

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

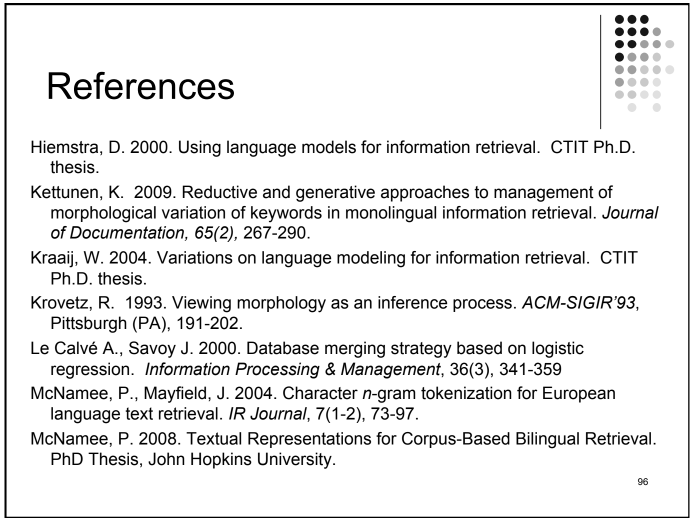
Gale, W.A., Church, K.W. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1), 75-102.

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.



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.

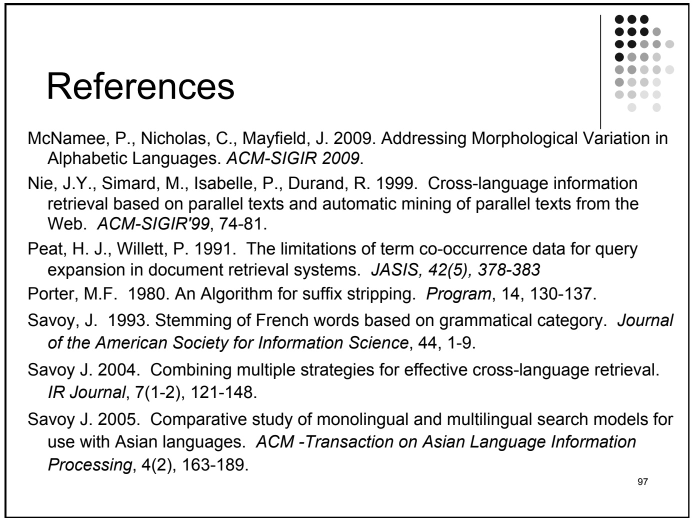
Kraaij, W. 2004. Variations on language modeling for information retrieval. CTIT Ph.D. thesis.

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.



References

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

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.

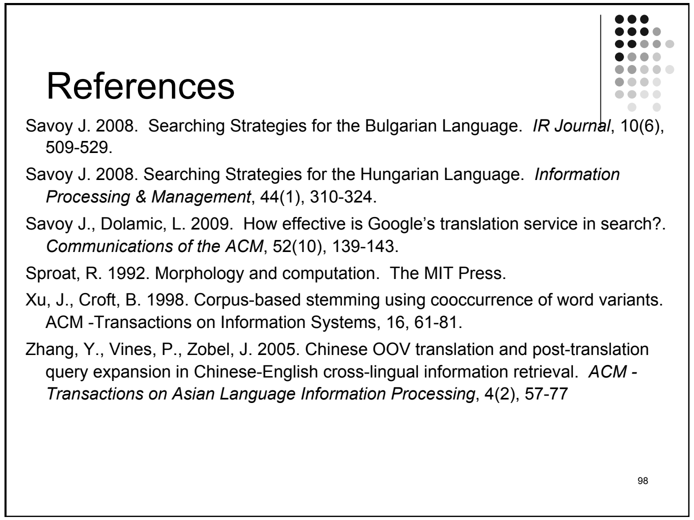
Peat, H. J., Willett, P. 1991. The limitations of term co-occurrence data for query expansion in document retrieval systems. *JASIS, 42(5), 378-383*

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

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

Savoy J., Dolamic, L. 2009. How effective is Google's translation service in search?. *Communications of the ACM*, 52(10), 139-143.

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.

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