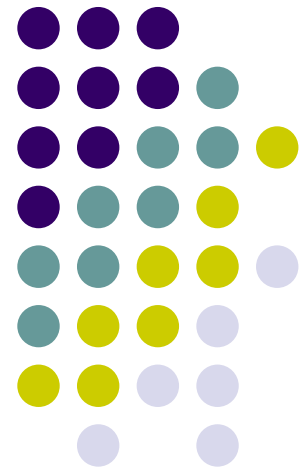
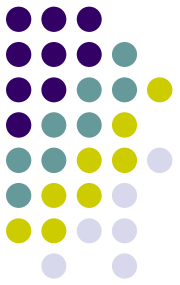


Searching across Different Languages

J. Savoy
University of Neuchatel

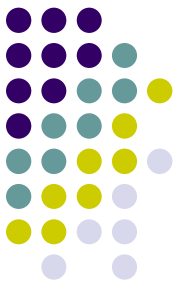
Various corpora cover a wide variety of natural languages, and search processes about a given item may be issued by users whose languages are extremely different.





Outline

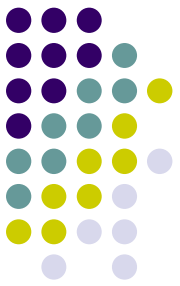
- Beyond just English: **Monolingual IR**
(encoding, diacritics, tokenization, segmentation, stopword list, stemming)
- Translation problems & strategies:
Bilingual IR



Beyond just English

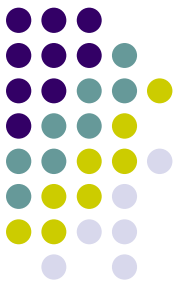
- Alphabets
 - Latin alphabet (26), Cyrillic (~33), Arabic (28), Hebrew, Hindi, Thai
- Ideograms
 - China (13,000/7,700) 中国人
 - Japan (8,800) 日本人
- Encoding systems
 - ASCII (limited to 7 bits)
 - **Windows-1251** (Cyrillic), BIG5, GB, EUC-JP, EUC-KR ...
 - ISO 8859-2 (East European), Cyrillic (**ISO-8859-5**), ...
 - Unicode (**UTF-8**, UTF-16, ...)

Monolingual IR

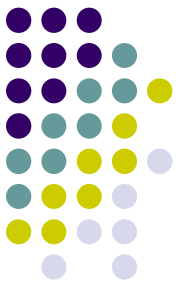


- Diacritics
 - differ from one language to another (“résumé”, “Äpfel”, “leão”)
 - could be used to distinguish the meaning (usually related)
 - usually there are removed by the IR system
(differences in MAP are usually small and non significant)
- Proper names
 - homophones involving proper names. E.g., 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)

Monolingual IR (Segmentation)

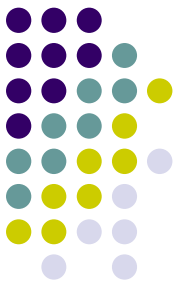


- What is a word / token?
 - Compound construction (worldwide, handgun) is used frequently in other languages (DE, NL, FI, HU)
 - In DE: “Bundesbankpräsident” =
“Bund” + es + “Bank” + “Präsident”
federal bank CEO
 - Important in DE: “Computersicherheit”
could appear as “die Sicherheit mit Computern”
 - Automatic decompounding in DE is useful
(+23% in MAP, short queries, +11% longer queries) [Braschler & Ripplinger 2004].
- Stopword list
 - Frequent and insignificant terms (+ pronouns, prep., conj.)
 - Differences in MAP with and without stopword list are usually small (non significant)



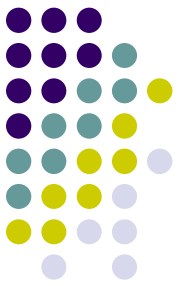
Monolingual IR (Stemming)

- Stemming (stems + rules = vocabulary)
 - Inflectional [Harman 1991]
 - the number (sing / plural) “horses” → “horse”
 - the gender (femi / masc / neutral)
 - verbal form (person, tense) “jumping” → “jump”Relatively simple in English (‘-s’, ‘-ing’, ‘-ed’)
 - Derivational [Porter 1980]
 - forming new words (changing POS)
 - ‘-ize’, ‘-ation’, ‘-ship’
 - compute → {computer, computing, computerize, computerization, computery}



Monolingual IR (Stemming)

- Rule-based (based on the grammar)
 - concentrate on the suffixes (ignore prefixes)
 - add quantitative constraints “king” → “k”?
 - add qualitative constraints
 - rewriting rules “running” → “runn” → “run”
- Could be adapted for specific domain (medicine)
- Over-stemming or under-stemming are possible (Porter)
 - “organization” → “organ”
 - “European” and “Europe” do not conflate
- Corpus-based (language usage) stemmer [Xu & Croft 1998]
- Using a dictionary (to reduce the error rate)
[Krovetz 1993], [Savoy 1993], [Hedlund et al. 2004]



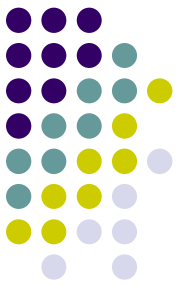
Monolingual IR (Slavic)

- Bulgarian (9M, Southern Slavic language, CLEF 2005-07)
 - Cyrillic alphabet
 - Three genders, two numbers (sing, plur)
 - No grammatical case (except vocative and pronouns)
 - Definite article
- Czech (11M, Western Slavic language, CLEF 2007)
 - Latin alphabet
 - Three genders (masc, femi, neutral), two numbers
 - Seven grammatical cases (also for names)
- Russian (165M, Eastern Slavic language, CLEF 2002-08)
 - Cyrillic alphabet
 - Three genders, two numbers
 - Six grammatical cases (also for names)
- Stemmers freely available at www.unine.ch/info/



Monolingual IR (Bulgarian)

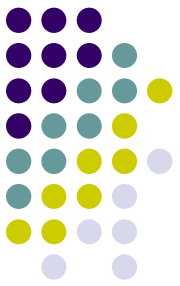
- Inflectional (gender, number, definite article)
 - слаб weak (masc, sing)
 - слаба (femi, sing)
 - слабата (femi, sing + the)
- Inflectional & derivationals
 - българ « stem »
 - България Bulgaria (noun)
 - българин Bulgarian (noun, masc, sing)
 - българка Bulgarian (noun, femi, sing)
 - българи Bulgarians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



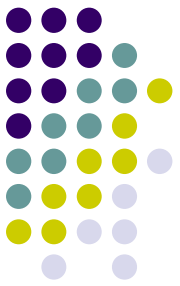
Monolingual IR (Bulgarian)

- 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, з, с
 - вълк → вълци (wolf → wolves)
 - герой → героят (hero → heros)

Monolingual IR (Czech)



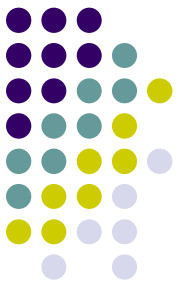
case gender	nominative	dative singulier	dative plural
Masculine (sir)	pán	pán <u>ovi</u>	pán <u>ům</u>
Feminine (woman)	žena <u>a</u>	žen <u>ě</u>	žen <u>ám</u>
Neutre (young)	mlad <u>é</u>	mlad <u>ému</u>	mlad <u>ým</u>



Monolingual IR (Czech)

- Even for names

Case \	Paris	Prague	France	Ann
nominative	Paříž	Praha	Francie	Anna
genitive	Paříž <u>e</u>	Prah <u>y</u>	Francie	An <u>ny</u>
dative	Paříž <u>í</u>	<i>Praze</i> <u>e</u>	Franci <u>í</u>	An <u>ě</u>
accusative	Paříž	Prah <u>u</u>	Franci <u>í</u>	An <u>u</u>
vocative	Paříž <u>í</u>	Prah <u>o</u>	Franci <u>e</u>	An <u>o</u>
locative	Paříž <u>í</u>	<i>Praze</i> <u>e</u>	Franci <u>í</u>	An <u>ě</u>
instrumental	Paříž <u>í</u>	Prah <u>ou</u>	Franci <u>í</u>	An <u>ou</u>



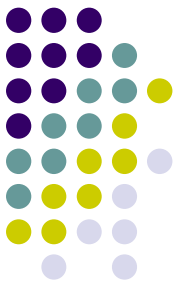
Monolingual IR (Czech)

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

Monolingual IR (Russian)



case gender	nominative	dative singular	dative plural
Masc. hard (city)	город	городу	город <u>ам</u>
Masc. soft (husband)	муж	мужу	муж <u>ьям</u>
Feminine (hand)	рука	руке	рук <u>ам</u>



Monolingual IR (Slavic)

- 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 (spelling variations, phonetic transliteration)
 - disc (EN) → “disk” (e.g., CZ)
 - “disc” (using the Latin alphabet)
 - “диск” (in Russian, Cyrillic alphabet)
 - Renault (EN) → “Renault” (e.g., CZ)
 - “Рено” (in Russian, Cyrillic alphabet)



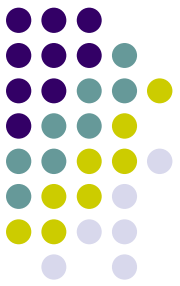
Monolingual IR (Bulgarian)

Stemming strategies, Bulgarian language [Savoy, 2008a]
Based on CLEF-2006-07 corpus, 99 TD queries

MAP	none	UniNE	Nakov
Okapi	0.2115	0.2805*	0.2642*
<i>tf·idf</i>	0.1697	0.1937*	0.2013*

Stopword list

MAP	none	UniNE (258 words)	BTB (804)
Okapi	0.2739	0.2805	0.2796
<i>tf·idf</i>	0.1928	0.1937	0.1930



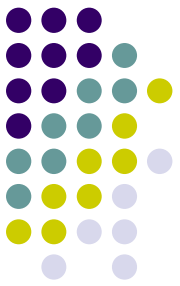
Monolingual IR (Czech)

Stemming strategies, Czech language

Based on CLEF-2008 corpus, 50 T queries

		UniNE	
MAP	none	light	aggres.
Okapi	0.2040	0.2990*	0.3065*
<i>tf·idf</i>	0.1357	0.2040*	0.2095*

With and without stopword list (467 words)
performance differences around 1%

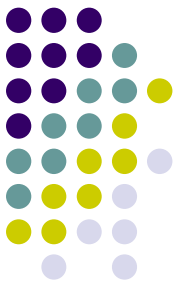


Monolingual IR (Hungarian-Finnish)

- Finno-Hungarian family owns numerous cases (HU, 18 cases [Savoy, 2008b])

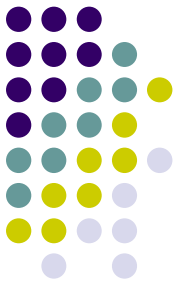
ház	nominative (house)
házat	accusative singular
házakat	accusative plural
házzal	“with” (instrumental)
házon	“over” (superessive)
házamat	my + accusative sing.
házamait	my + accusative + plur.

- In FI, the stem may change! (more regular in HU) (e.g., “matto”, “maton”, “mattoja” (carpet))
Need a deeper morphological analyzer is useful for FI (see [Tomlinson, 2005], [Hedlund et al. 2004])
- Compound construction possible
HU: “internetfüggők” → “internet” + “függők” (+ addiction)
FI: “rakkauskirje” → “rakkaus” + “kirje” (love + letter)



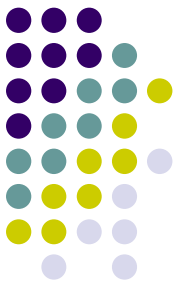
Monolingual IR (Stemming)

- Mean relative improvement (MAP) after stemming
 - +4% English
 - +4% Dutch
 - +7% Spanish
 - +9% French
 - +15% Italian
 - +19% German
 - +29% Swedish
 - +34% Bulgarian
 - +40% Finnish
 - +44% Czech
- Evaluations: CLEF proceedings
 - Stemming > none
 - Differences between stemmers could be statistically significant
 - Simple stemmers for nouns & adjectives tend to perform better, or at the same level of performance than more aggressive stemmers



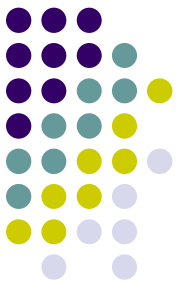
Outline

- Beyond just English: Monolingual IR
(encoding, diacritics, tokenization,
segmentation, stopword list, stemming)
- Translation problems & strategies:
Bilingual IR



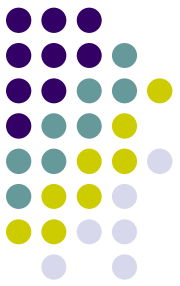
Translation Ambiguity

- “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)
- “light” (EN → FR), “lumière” or “clair” “léger” POS may help
 - As noun → “lumière”
 - As adjective → “clair”, “léger”



Translation Tools

- Machine-readable bilingual dictionaries (MRD)
 - provide usually more than one translation alternatives (take all? the first?, same weight for all?)
 - OOV problem (e.g., proper nouns)
 - Could be limited to simple word lists
 - Must provide the *lemmas* (not the surface words!)
- Machine Translation system (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 & included inside the IR model



Translation (Google MT)

Search into an EN corpus, queries (284) (T-only)

Automatic translation done by Google's MT (May 2007)

Statistical significant difference (*) [Dolamic & Savoy 2009]

MAP	Mono	From ZH	From DE	From FR	From SP
Okapi	0.4044	0.3327*	0.3625*	0.3692*	0.3752*
LM	0.3708	0.3019*	0.3305*	0.3400*	0.3426*
<i>tf idf</i>	0.2392	0.1920*	0.2266*	0.2294*	0.2256*
<i>diff</i>		-18.2%	-9.3%	-7.3%	-7.1%

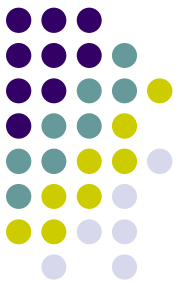


Translation (Error Distribution)

Where are the real translation problems?

For Google's 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



Translation (Pivot Language)

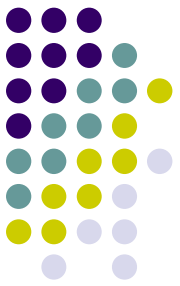
Search into FR corpus, 299 queries (CLEF 2001-06)

Original queries in FR (Title-only) [Savoy & Dolamic 2009]

Query language is German

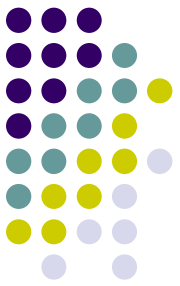
MRR	Mono	From EN	From DE	From DE-EN
Okapi	0.6631	0.5817	0.4631	0.5273
Diff.		-12.3%	-30.2%	-20.5%

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



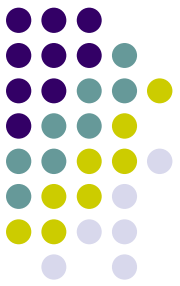
Conclusion

- Monolingual
 - could be relatively simple for foreign languages close to English (Romance, Germanic, and even Slavic families)
 - Diacritics & stopword list: not a real problem
 - Compound construction is important DE
 - More morphological analysis could clearly improved the IR performance (FI)
 - Some test-collections are problematic (AR in TREC 2001, RU in CLEF 2004)
- 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 (using MRR)



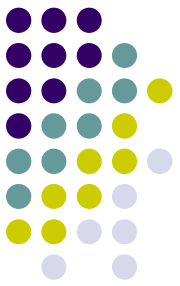
Conclusion

- 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 (using MRR)
 - merging is not fully resolved (e.g. using ML approach? see [Si & Callan, 2006])
- <http://www.clef-campaign.org>
- <http://research.nii.ac.jp/ntcir/>
- <http://trec.nist.gov> (TREC-3 to TREC-12)
- <http://romip.ru/en> (in Russian language only)



References

- Braschler, M., Ripplinger, B. 2004. How effective is stemming and compounding for German text retrieval? *IR Journal*, 7, 291-316.
- Braschler, M. 2004. Combination approaches for multilingual text retrieval. *Information Retrieval*, 7(1-2), 183-204.
- Dolamic, L., Savoy, J. 2009. Monolingual and bilingual searches: Evaluation, challenges and failure analysis. Submitted.
- Gao, J., Nie, J.-Y. 2006. A study of statistical models for query translation: Finding a good unit of translation. *ACM-SIGIR'2006*. Seattle (WA), 194-201.
- 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), 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.
- Krovetz, R. 1993. Viewing morphology as an inference process. *ACM-SIGIR'93*. Pittsburgh (PA), 191-202.



References

- Moulinier, I. 2004. Thomson Legal and Regulatory at NTCIR-4: Monolingual and pivot-language retrieval experiments. NTCIR-4 Workshop, Tokyo, 1-8
- Porter, M.F. 1980. An Algorithm for suffix stripping. Program, 14, 130-137.
- Savoy, J. 1993. Stemming of French words based on grammatical category. Journal of the American Society for Information Science, 44, 1-9.
- Savoy J. 2004. Combining multiple strategies for effective cross-language retrieval. IR Journal, 7(1-2), 121-148.
- Savoy J. 2005. Comparative study of monolingual and multilingual search models for use with Asian languages. ACM -Transaction on Asian Language Information Processing, 4(2), 163-189.
- Savoy J. 2008a. Searching Strategies for the Bulgarian Language. IR Journal, 10(6), 2008, 509-529.
- Savoy J. 2008b. 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.
- Tomlinson, S. 2005. Finnish, Portuguese and Russian with Hummingbird SearchServer™ at CLEF2004. C. Peters et al. Multilingual Information Access for Text, Speech and Images, LNCS #3491, Springer, 221-232.
- Xu, J., Croft, B. 1998. Corpus-based stemming using cooccurrence of word variants. ACM -Transactions on Information Systems, 16, 61-81.