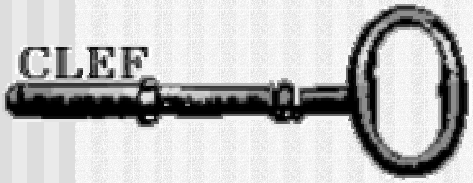
Multilingual Search using Query Translation and Collection Selection



Jacques Savoy, Pierre-Yves Berger
University of Neuchatel, Switzerland
www.unine.ch/info/clef/

Our IR Strategy

1. Effective monolingual IR system (combining indexing schemes?)
2. Combining query translation tools
3. Effective collection selection and merging strategies

Effective Monolingual IR

1. Define a stopword list
2. Have a « good » stemmer

   Removing only the feminine and plural suffixes (inflections)
   Focus on nouns and adjectives
   see www.unine.ch/info/clef/

Indexing Units

For Indo-European languages :
Set of significant words

For Finnish, we used 4-grams.

CJK: bigrams
with stoplist
without stemming

IR Models

Probabilistic

- Okapi
- Prosit or deviation from randomness

Vector-space

- Lnu-ltc
- tf-idf (ntc-ntc)
- binary (bnn-bnn)

Monolingual Evaluation

Model	English	French	Portug.
Q=TD	word	word	word
Okapi	**0.5422**	**0.4685**	**0.4835**
Prosit	0.5313	0.4568	0.4695
Lnu-ltc	0.4979	0.4349	0.4579
tf-idf	0.3706	0.3309	0.3708
binary	0.3005	0.2017	0.1834

Monolingual Evaluation

Model	Finnish		Russian	
Q=TD	word		word	
Okapi	0.4773		**0.3800**	
Prosit	0.4620		0.3448	
Lnu-ltc	0.4643		0.3794	
tf-idf	0.3862		0.2716	
binary	0.1859		0.1512	

Monolingual Evaluation

Model	Finnish		Russian	
Q=TD	word	**4-gram**	**word**	4-gram
Okapi	0.4773	**0.5386**	**0.3800**	0.2890
Prosit	0.4620	0.5357	0.3448	0.2879
Lnu-ltc	0.4643	0.5022	0.3794	0.2852
tf-idf	0.3862	0.4466	0.2716	0.1916
binary	0.1859	0.2387	0.1512	0.0373

Problems with Finnish

Finnish can be characterized by:

1. A large number of cases (12+),
2. Agglutinative (sauna**stani**)
   & compound construction (rakkauskirje)
3. Variations in the stem
   (matto, mato**n**, mattoj**a**, …).

Do we need a deeper morpholgical analysis?

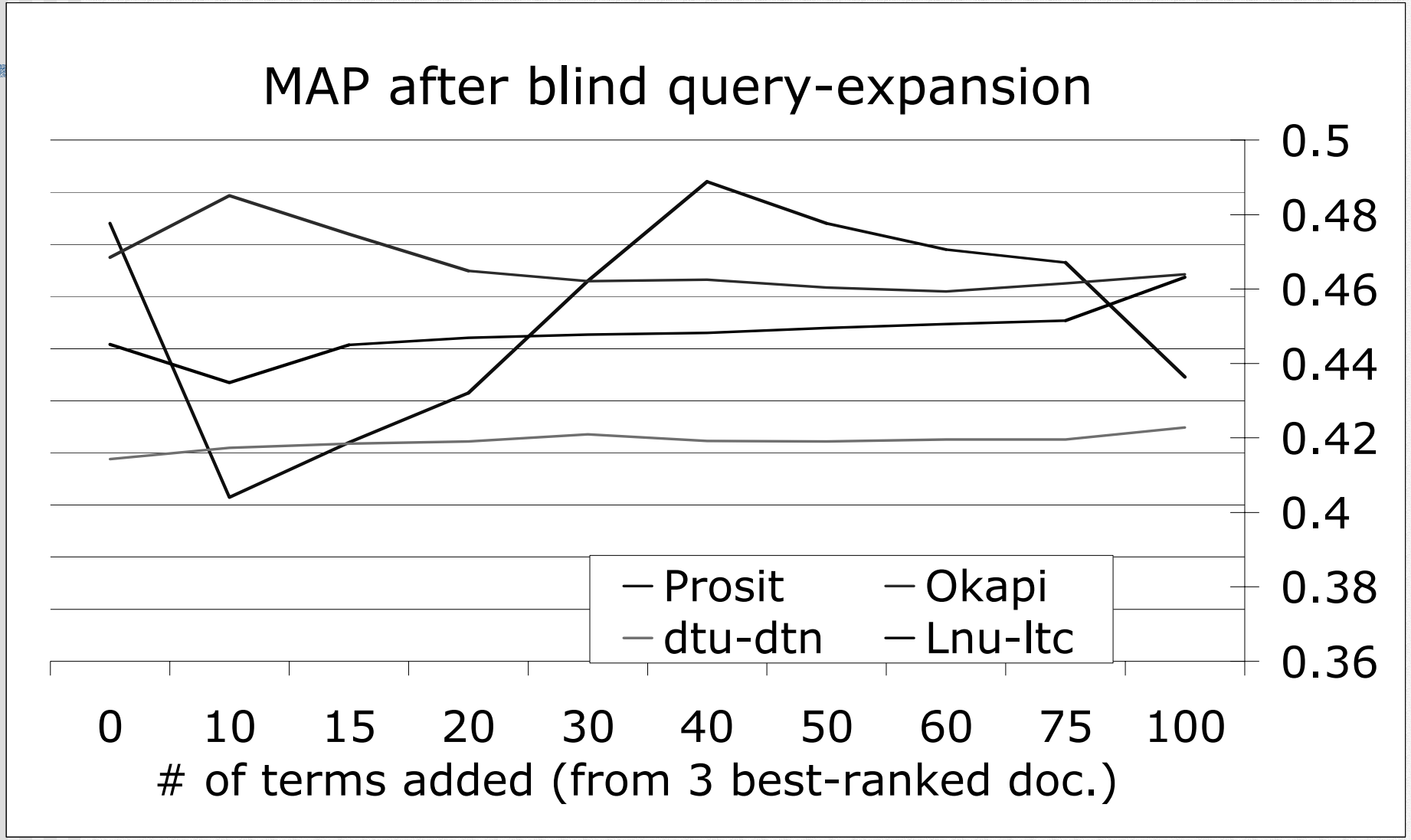
Improvement using…

Relevance feedback (Rocchio)?

Monolingual Evaluation

Q=TD	English	French	Finnish	Russian
Okapi	0.5422	0.4685	0.5386	0.3800
+PRF	0.5704	**0.4851**	0.5308	**0.3945**
	+5.2%	+3.5%	-1.5%	+3.8%
Prosit	0.5313	0.4568	0.5357	0.3448
+PRF	**0.5742**	0.4643	**0.5684**	0.3736
	+8.1%	+1.6%	+6.1%	+8.4%

Relevance Feedback



Improvement using…

Data fusion approaches (see the paper)

Improvement with the English, Finnish and Portuguese corpora

Hurt the MAP with the French and Russian collections

Bilingual IR

1. Machine-readable dictionaries (MRD)
   e.g., Babylon
2. Machine translation services (MT)
   e.g., FreeTranslation
   BabelFish
3. Parallel and/or comparable corpora
   (not used in this evaluation campaign)

Bilingual IR

"Tour de France Winner
Who won the Tour de France in 1995?"

(Babylon1) "France, vainqueur
Organisation Mondiale de la Santé, le, France,* 1995"

(Right) "Le vainqueur du Tour de France
Qui a gagné le tour de France en 1995 ?"

Bilingual EN->FR/FI/RU/PT

TD Okapi	French word	Finnish 4-gram	Russian word	Portug word
Manual	0.4685	0.5386	0.3800	0.4835
Baby1	0.3706	0.1965	0.2209	0.3071
FreeTra	0.3845	N/A	0.3067	0.4057
InterTra	0.2664	0.2653	0.1216	0.3277
Reverso	0.3830	N/A	0.2960	N/A
Combi.	**0.4066**	**0.3042**	**0.3888**	**0.4204**

Bilingual EN->FR/FI/RU/PT

TD Okapi	French word	Finnish 4-gram	Russian word	Portug word
Manual	0.4685	0.5386	0.3800	0.4835
Baby1	79.1%	36.5%	58.1%	63.5%
FreeTra	82.1%	N/A	80.7%	83.9%
InterTra	56.9%	49.3%	32.0%	67.8%
Reverso	81.8%	N/A	77.9%	N/A
Combi.	**86.8%**	**56.5%**	**102.3%**	**86.9%**

Bilingual IR

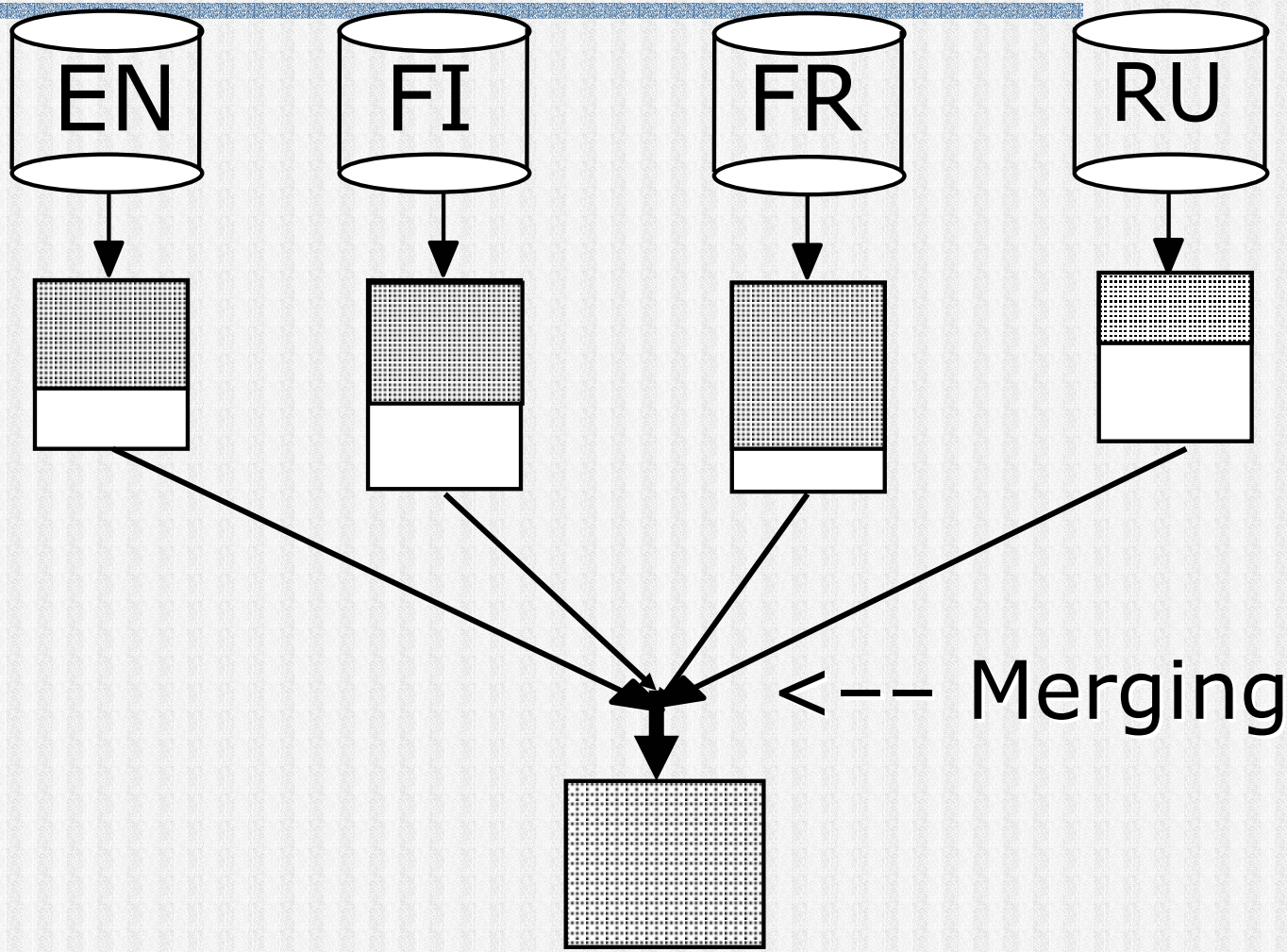
Improvement …

- Query combination (concatenation)
- Relevance feedback
  (yes for PT, FR, RU, no for FI)
- Data fusion
  (yes for FR, no for RU, FI, PT)

Multilingual EN->EN FI FR RU

1. Create a common index
   Document translation (DT)
2. Search on each language and
   merge the result lists (QT)
3. Mix QT and DT
4. No translation

Merging Problem



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	1.6
2	RU005	1.1
3	RU120	0.9
4	…	

Merging Strategies

- Round-robin (baseline)
- Raw-score merging
- Normalize
- Biased round-robin
- Z-score
- Logistic regression

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	1.6
2	RU005	1.1
3	RU120	0.9
4	…	

1 RU…
2 EN…
3 RU…
4 ….

Merging Strategies

- Round-robin (baseline)
- Raw-score merging
- Normalize
- **Biased round-robin**
- Z-score
- Logistic regression

Test-Collection CLEF-2004

	EN	FI	FR	RU
size	154 MB	137 MB	**244** MB	**68** MB
doc	56,472	55,344	90,261	16,716
topic	42	45	49	34
rel./query	8.9	9.2	**18.7**	**3.6**

Merging Strategies

- Round-robin (baseline)
- Raw-score merging
- Normalize
- Biased round-robin
- **Z-score**
- Logistic regression

Z-Score Normalization

1	EN120	1.2
2	EN200	1.0
3	EN050	0.7
4	EN765	0.6
…		…

Compute the mean μ and standard deviation σ

New score =
$((\text{old score} - \mu) / \sigma) + \delta$

1	EN120	7.0
2	EN200	5.0
3	EN050	2.0
4	EN765	1.0
…		

MLIR Evaluation

Condition A (meta-search)

Prosit/Okapi (+PRF, data fusion for FR, EN)

Condition C (same SE)

Prosit only (+PRF)

Multilingual Evaluation

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

Collection Selection

General idea:

1. Use the logistic regression to compute a probability of relevance for each document;
2. *Sum* the document scores from the first 15 best ranked documents;
3. If this *sum* $\geq \delta$, consider all the list, otherwise select only the first *m* doc.

Multilingual Evaluation

EN->EN FR FI RU	Cond. A	Cond. C
Round-robin	0.2386	0.2358
Logistic reg.	0.3090	0.3393
Optimal select	**0.3234**	**0.3558**
Select (m=0)	0.2957	0.3405
Select (m=3)	0.2953	0.3378

Conclusion (monolingual)

- Finnish is still a difficult language
- Blind-query expansion:
  different patterns with
  different IR models
- Data fusion (?)

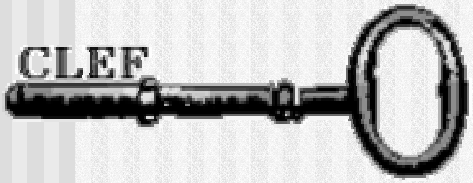
Conclusion (bilingual)

- Translation resources not available for all languages pairs
- Improvement by combining query translations yes, but can we have a better combination scheme?

Conclusion (multilingual)

- Selection and merging are still hard problems
- Overfitting is not always the best choice (see results with Condition C)

Multilingual Search using Query Translation and Collection Selection



Jacques Savoy, Pierre-Yves Berger
University of Neuchatel, Switzerland
www.unine.ch/info/clef/