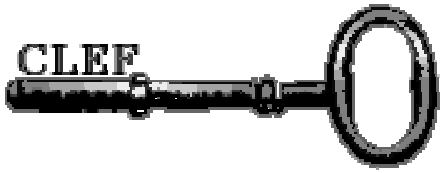
Combined strategies for effective multilingual IR



Jacques Savoy

University of Neuchatel, Switzerland

www.unine.ch/info/clef/

Important features for MLIR

- Effective monolingual IR system (combining indexing schemes)
- Combining query translation tools
- Effective collection fusion strategy

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
available at www.unine.ch/info/clef/

Monolingual IR

Improvements in 2002

Extended stoplist in French and German

Some derivational suffixes in French

Data fusion for German collection

Okapi is still the best approach

Monolingual IR

For French, Italian, Spanish, English

we used whole words

For German, Dutch & Finnish

combining word, 5-gram, decompounding

(McNamee & Mayfield, CLEF-2001)

Monolingual IR

5-gram indexing

"das Hausdach" -> "das", "Hausd", "ausda"

Decompounding

Set of patterns, e.g. "-ung -ung -"

"Betreuungsstelle" ->

"Betreuung", "Stelle", ...

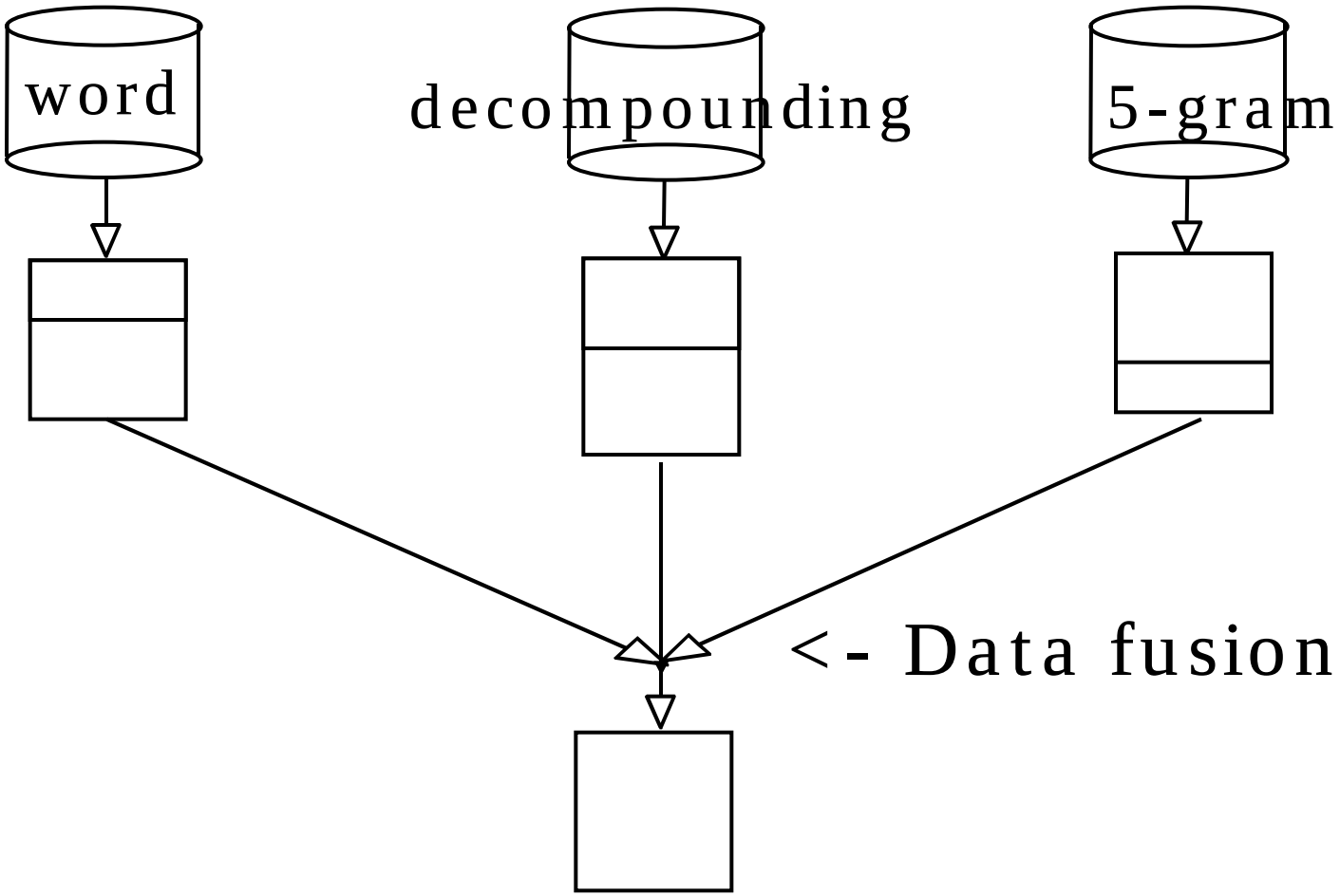
Monolingual IR

How can we combine these runs?

Query=TD

Okapi	words	decomp.	5-gram
German	37.4	37.8	39.8
Dutch	42.4		41.8
Finnish	31.0		38.3

Data fusion strategies



Data fusion strategies

- Round robin (Voorhees et al., TREC'3)
- combSUM (Fox and Shaw, TREC'2)

  SUM(RSV_i)
- MAX Max(RSV_i)
- combNBZ SUM(RSV_i) * # nonzero
- Logistic regression (Le Calvé et al., IPM, 2000)
- CORI (Callan, SIGIR'02)

Round Robin

1	GE120	1.2
2	GE200	1.0
3	GE050	0.7
4	GE765	0.6
...		
8	GE567	0.2

1	GE043	0.8
2	GE120	0.75
3	GE055	0.65

1	GE050	1.6
2	GE195	1.3
3	GE120	0.9
4	GE649	0.7
...		
12	GE200	0.1

1	GE120
2	GE043
3	GE050
4	GE200
5	GE055
...	

combSUM

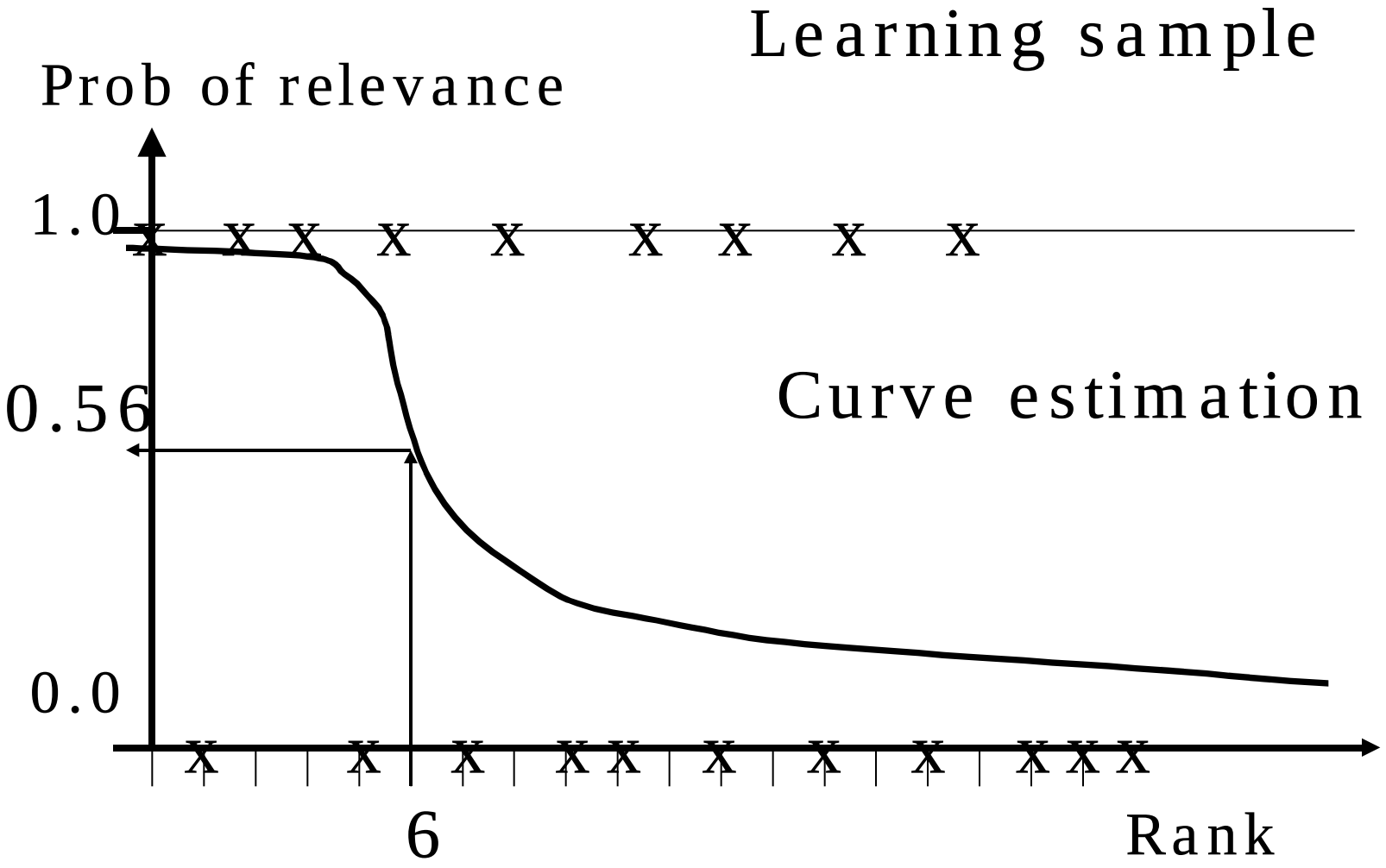
1	GE120	1.2
2	GE200	1.0
3	GE050	0.7
4	GE765	0.6
...		
8	GE567	0.2

1	GE043	0.8
2	GE120	0.75
3	GE055	0.65

1	GE050	1.6
2	GE195	1.3
3	GE120	0.9
4	GE649	0.7
...		
12	GE200	0.1

1	GE120	2.85
2	GE050	2.3
3	GE195	1.3
4	GE200	1.1
5	GE043	0.8
...		

Logistic Regression



Monolingual IR

Query=TD	German words	German decomp.	German 5-gram
Okapi	37.4	37.8	39.8
Round robin		40.2 (+0.9%)	
combSUM		**42.3** (+6.2%)	
Logistic regr.		42.0 (+5.4%)	
combNBZ		41.5 (+4.2%)	
CORI		41.3 (+3.6%)	

CLEF-01 vs. CLEF-02

CLEF-02 collection, Query=TD, Okapi

Setting	CLEF-01	CLEF-02	
German	38.3	39.5	(+3.3%)
combSUM		42.3	(+11.6%)

Query translation tools

We used

- Machine translation tools (5)
  - BabelFish
  - Reverso
  - FreeTranslation
  - ...
- Bilingual dictionary
  - www.babylon.com

Query translation

«AI in Latin America»

Reverso -> «AI en Amérique latine»
Google -> «AI en Amérique latine»
InterTrans -> «AI dans Amérique latine»
Babylon -> «intelligence artificielle dans latin Amérique»

Query translation

«AI in Latin America»

As a combined query
«AI en Amérique latine
AI en Amérique latine
AI dans Amérique latine
intelligence artificielle
dans latin Amérique»

Query translation

Query = TD, Okapi

	French	Spanish	German 5-gram
Manually	48.4	51.7	39.8
Baylon 1	**43.2**	39.6	28.1
Systran	42.7	38.5	27.7
Reverso	39.0	**43.3**	**28.7**

Query translation

Query = TD, Okapi

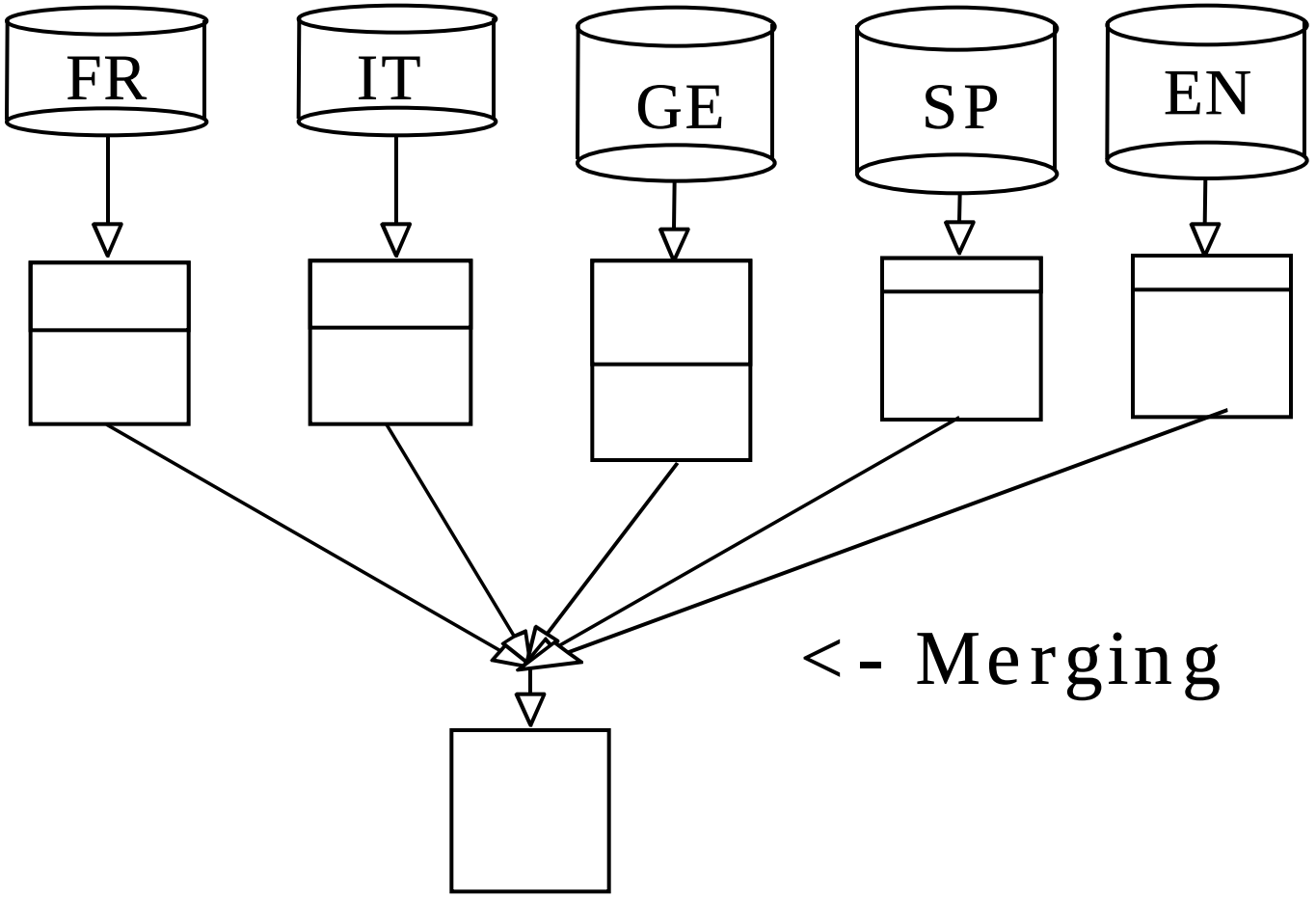
	French	Spanish	German 5-gram
Manually	48.4	51.7	39.8
Best single	43.2	43.3	28.7
Best Comb. QT	48.6	45.6	33.3
+ data fusion			38.7

CLEF-01 vs. CLEF02

CLEF-02 collection, Query=TD, Okapi

Bilingual E->	French	Spanish	German 5-gram
CLEF-01	46.6	44.0	31.9
CLEF-02	48.6	45.6	33.3
	+4.1%	**+3.8%**	**+4.6%**
+ data fusion			38.7
			+21.5%

Collection fusion strategies



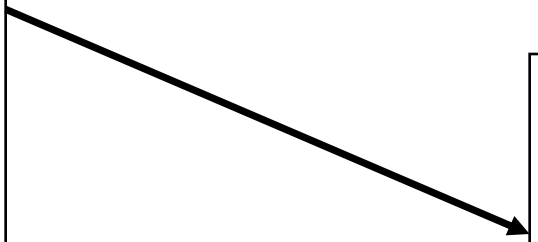
Collection fusion strategies

- Round robin
- combRSV% (normalized score)
- CORI
- Logistic regression

combRSV%

1	IT123	1.2
2	IT673	1.0
3	IT946	0.72
4	IT765	0.6
...		
8	IT567	0.2

divided by the max score



1	IT123	1.0
2	IT673	0.833
3	IT946	0.6
4	IT765	0.5
...		
8	IT567	0.166

sort the result lists using this new score

Collection fusion strategies

Query = TD

Okapi	manually	automatic
- round robin	33.9	31.2
- combRSV%	35.1 (+4%)	31.8 (+2%)
- CORI	36.4 (+8%)	34.0 (+9%)
- Logistic regr.	**38.8** (+15%)	**34.9** (+12%)

Collection fusion strategies

Query = TD and blind query expansion

Okapi	manually	automatic
- round robin	36.8	34.6
- combRSV%	35.1 (-2%)	33.8 (-2%)
- CORI	36.4 (+2%)	36.8 (+6%)
- Logistic regr.	**43.8** (+17%)	**39.8** (+15%)

without blind query expansion

round robin	33.9	31.2

Errare humanum est

Query #130 and Query #131 have 0.00 average precision in our official bilingual and multilingual runs ...

Switching these two queries in the English topics set.

Errare humanum est

Official run	37.83
Corrected run	39.49 !

Yes, but with the logistic regression...
(thus with a learning stage)

Berkeley 2	37.62
Neuchatel	36.62 UniNEm1

learning improves by +7.8%

Conclusion
CLEF-01 vs. CLEF-02

Strategies	2001	2002
- German	38.3	42.3
(data fusion)		+10.6%
- Bilingual Italian	32.7	35.8
(multiple QT)		+9.6%
- MLIR	28.6	34.9
(without Qexp)		+22.1%