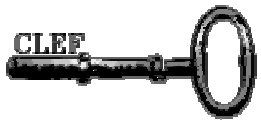
Report on CLEF-2001 Experiments

Jacques Savoy



University of Neuchatel (Switzerland)

www.unine.ch/info/clef/

Important features for MLIR

- Effective monolingual IR system
- Combining query translation tools
- Effective database merging strategy

Based on fully automatic approach

Monolingual IR

Define a stopword list

- Very frequent words (200)

  and we add:

- Articles (the, an, few, ...)
- Prepositions (in, at, ...)
- Conjunctions (like, as, ...)
- Auxiliaries (to be, has, do, ...)
- Pronouns (you, mine, whom, ...)

Monolingual IR

Have a «good» stemmer

Based on morphological decomposition
to remove «-ably», «-alistic», ...

Here, we only remove the feminine
and plural suffixes (inflections)

Focus on nouns and adjectives

Monolingual IR

Removing inflectional suffixes

masculine singular	el amig<u>o</u>
feminine singular	la amig<u>a</u>
masculine plural	los amig<u>os</u>
feminine plural	los amig<u>as</u>

available www.unine.ch/info/clef/

Monolingual IR

Inflection removals (only)

is effective in terms of precision/recall

Cheval / chevaux -> cheval (horse)
Chevalerie (knighthood) -> cheval
Chevalet (easel) -> cheval

suggested by www.Yahoo.fr

Monolingual IR

For French, Italian, Spanish, English

we used whole words

For German collection

we used 5-grams (effective CLEF2000)

(McNamee & Mayfield, CLEF-2000)

Monolingual IR

- binary {0, 1} (bnn)
- tf = occurrence frequency (nnn)
- idf = inverse document frequency
- tf · idf, (ntc)
- log(tf) · idf, (ltc)
- length of the document (Okapi)

10 different retrieval models

Monolingual IR

Query = TD

	French	German 5-grams	German words
Okapi	49.9	39.5	39.1
atn-ntc	48.0	37.5	35.8
Lnu-ltc	47.4	36.7	37.9
ntc-ntc	32.2	30.1	27.6
bnn-bnn	17.9	18.8	21.1
nnn-nnn	14.1	9.8	13.8

Monolingual IR

Based on five different languages

Baseline Q = Title only

	Q=TD	Q=TDN	Q=TDN full
Okapi	+15%	+21.2%	+28.3%
mean 10 IR	+13.3%	+17.5%	+34.8%

full: indexing documents with all logical sections

Query translation tools

We used

- Machine translation tool
  babel.altavista.com
- Bilingual dictionary
  www.babylon.com

Query translation tools

Based on four different languages

Baseline Q = TD, manually translated

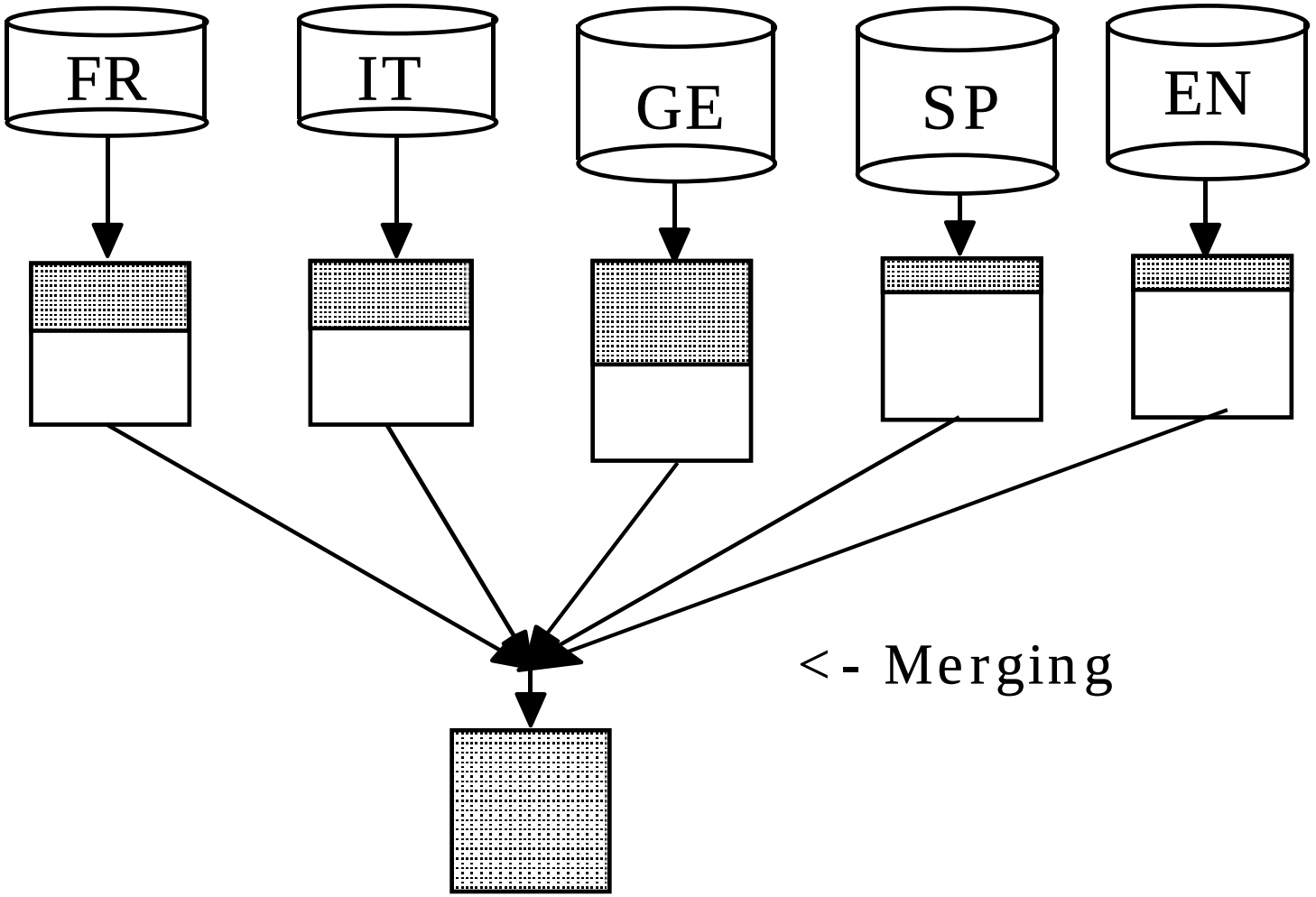
	Systran	Babylon	Combined
Okapi	-21.8%	-31.3%	-12.3%
mean 10 IR	-25.9%	-33.7%	-17.0%

Query translation tools

Failure analysis

- Too small coverage of the dictionary
  (e.g., Euthanasia)
- Ambiguity of the search term
- Incorrect identification of multi-words concept
  (e.g., renewable power)
- Translation of proper nouns (e.g., Latvia)

Database merging strategies



Database merging strategies

- Round robin (Voorhees et al., TREC'3)
- Raw score merging
- CORI (Callan et al., SIGIR'95)
- Normalized score

Round robin

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

1	FR453	0.8
2	FR012	0.75
3	FR673	0.65

1	GE567	1.6
2	GE195	1.3
3	GE548	0.9
4	GE649	0.7
...		
12	GE940	0.1

1	IT123
2	FR453
3	GE567
4	IT673
5	FR012
...	

Raw score merging

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

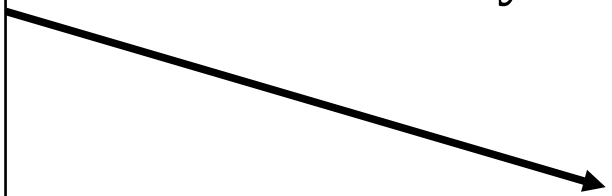
1	FR453	0.8
2	FR012	0.75
3	FR673	0.65

1	GE567	1.6
2	GE195	1.3
3	GE548	0.9
4	GE649	0.7
...		
12	GE940	0.1

1	GE567	1.6
2	GE195	1.3
3	IT123	1.2
4	IT673	1.0
5	GE548	0.9
6	FR453	0.8

Normalized score merging

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

Database merging strategies

Query = TD

Okapi	original	MLIR	change
- round robin	34.2	29.6	-13.6%
- raw score	15.9	13.1	-17.6%
- CORI	13.9	11.2	-19.2%
- normalized score	**38.0**	**31.3**	-17.8%

Database merging strategies

Failure analysis: percentage of documents

	Qrel	RR	CORI	Norm
English	10.5%	20%	0.3%	22.4%
French	14.9%	20%	0.9%	19.0%
Italian	15.3%	20%	0.4%	18.0%
German	26.2%	20%	**97.9%**	19.5%
Spanish	33.1%	20%	0.4%	21.2%

Database merging strategies

Query = TD & blind query expansion

Okapi	original	MLIR	change
- round robin	36.8	33.1	-10.2%
- raw score	6.1	6.1	0.0%
- normalized score	**41.1**	**34.3**	-16.6%

without blind query expansion

normalized score	38.0	31.3	

Conclusion

- Effective monolingual IR (Okapi, inflectional suffixes)
- Combined query translation tools
- Blind query expansion (CLIR)
- Result lists merging strategy (normalized score)