Cross-language Retrieval Experiments at CLEF-2002

Aitao Chen

School of Information Management and Systems

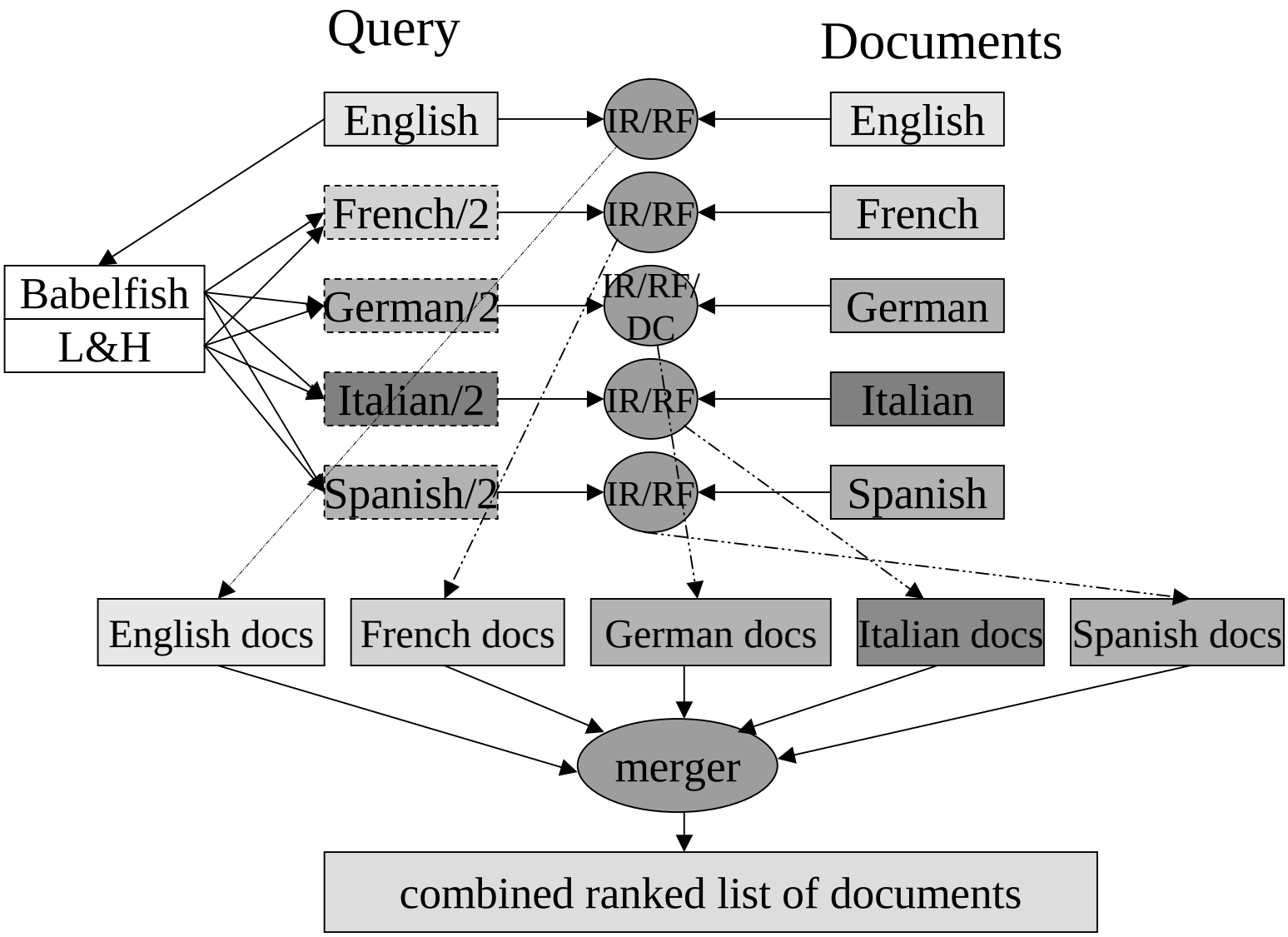
University of California at Berkeley

CLEF 2002 Workshop: 19-20 September, 2002, Rome, Italy

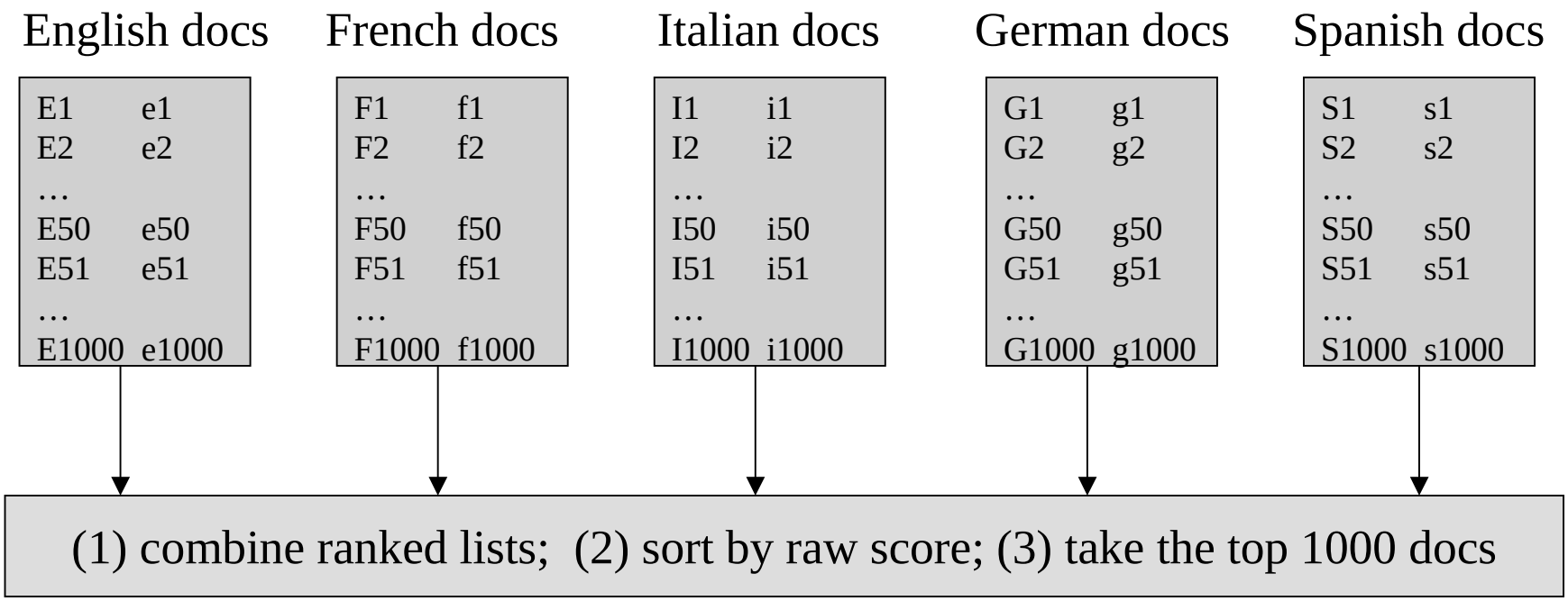
Talk Outline

- Overview of our CLEF-2002 experiments
- Evaluation of merging strategies
- German/Dutch decompounding
- Query expansion
- Conclusions

Overview of Multilingual Information Retrieval Experiments

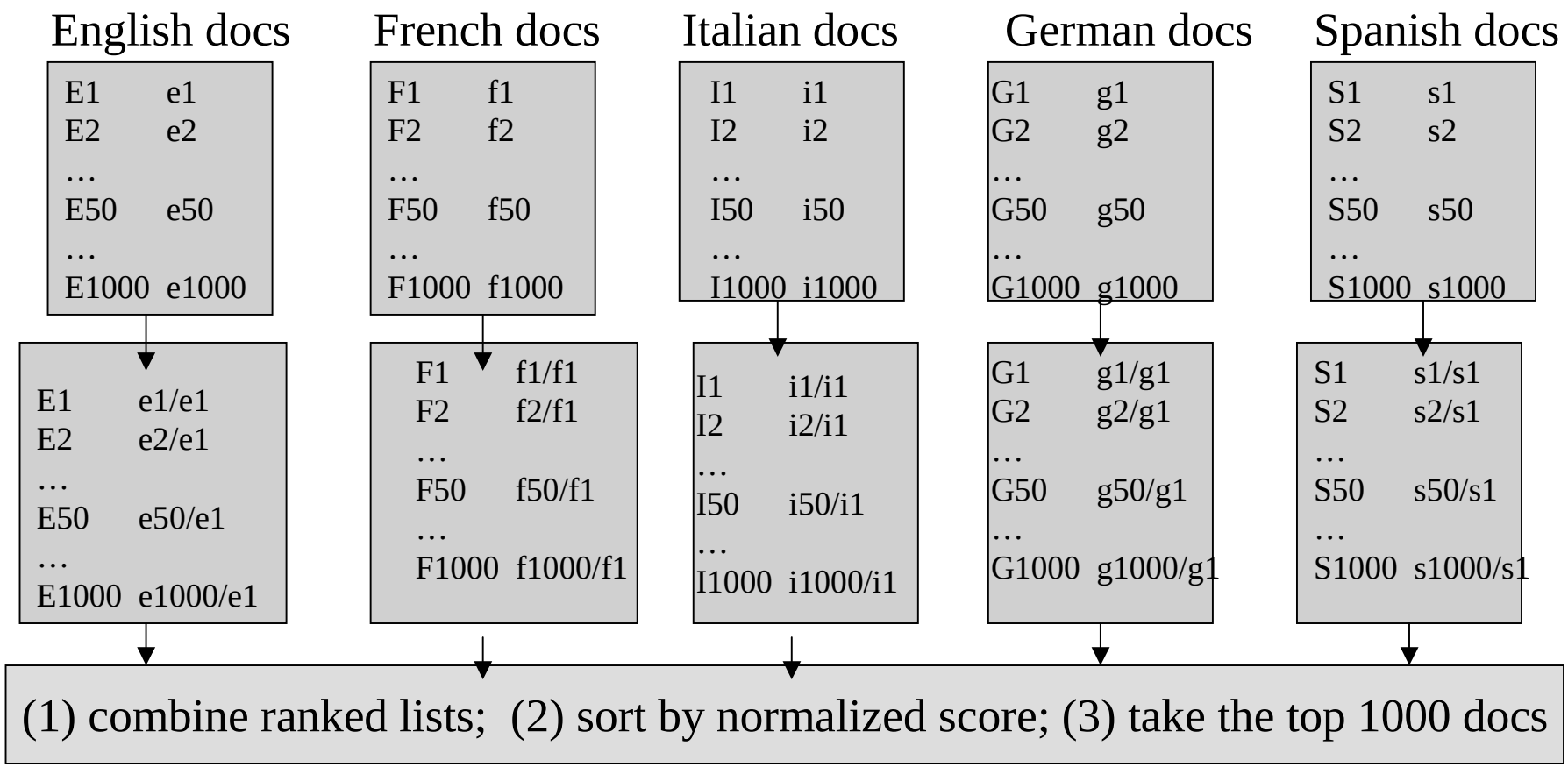


Multilingual Information Retrieval: Direct Merging



Weakness: prone to un-comparable raw relevance scores in individual ranked lists.

Multilingual Information Retrieval: Normalized Merging



Weakness: prone to skewed distribution of relevant documents over the document subcollections.

Optimal Merging (Known Relevance)

(red=rel doc; black=irrel doc)

Table 1

Rank	Run A	Run B	Run C
1	A1	B1	C1
2	A2	B2	C2
3	A3	B3	C3
4	A4	B4	C4

Table 2

Set	Run A	Run B	Run C
1	(0,1) {A1}	(2,1) {B1,B2,B3}	(1,3) {C1,C2,C3,C4}
2	(1,1) {A2,A3}	(1,0) {B4}	
3	(1,0) {A3}		

Table 3

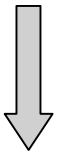
Set	Optimal Ranking
1	(0,1) {A1}
2	
3	
4	
5	
6	

Choose the set with the smallest number of irrelevant documents, but the largest number of relevant documents from the set of active sets.

Optimal Merging (Known Relevance)

(red=rel doc; black=irrel doc)

Rank	Run A	Run B	Run C
1	A1	B1	C1
2	A2	B2	C2
3	A3	B3	C3
4	A4	B4	C4

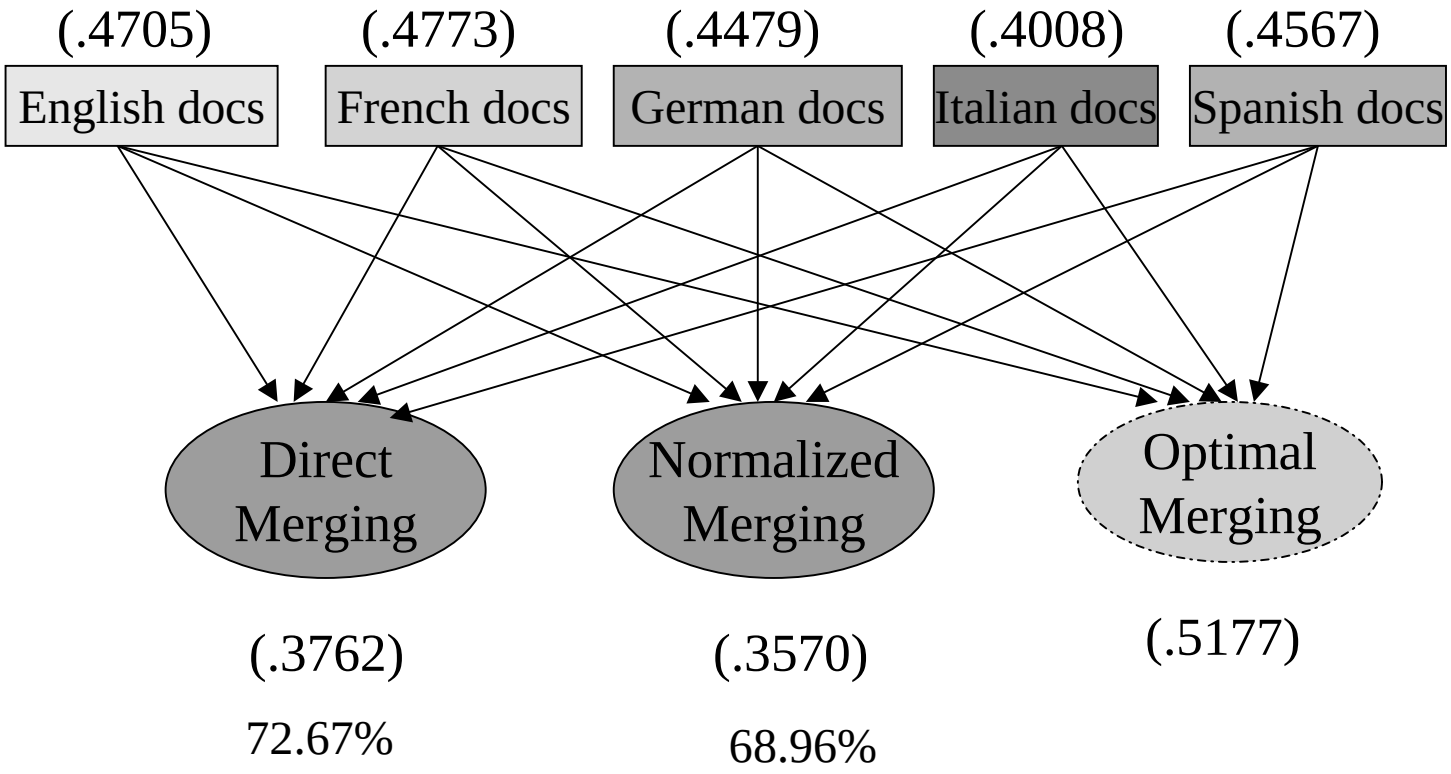


Set	Run A	Run B	Run C
1		(2,1) {B1,B2,B3}	(1,3) {C1,C2,C3,C4}
2	(1,1) {A2,A3}	(1,0) {B4}	
3	(1,0) {A3}		

Set	Optimal Ranking
1	(0,1) {A1}
2	(1,3) {C1,C2,C3,C4}
3	(1,1) {A2,A3}
4	(2,1) {B1,B2,B3}
5	(1,0) {A4}
6	(1,0) {B4}

1	A1
2	C1
3	C2
4	C3
5	C4
6	A2
7	A3
8	B1
9	B2
10	B3
11	A4
12	B4

Performances of MLIR with Different Merging Strategies (Topics: English,TD)



	English	French	German	Italian	Spanish
No. topics with no rel docs	8	0	0	1	0

Why decompounding?

- Topic 109: “Computersicherheit” (computer security) in title & desc fields, but not in the German document collection.
- Topic 88: The Dutch compound “gekkekoeienziekte” (mad cow disease) is not in the Dutch document collection.
- Topic 113: “Fussballeuropameisterschaft” in the title, but “Europameisterschaft im FuBball” in the desc. (B/ss)
- Topic 115: “Scheidungsstatistiken” (divorce statistics) in the title, but “Statistiken uber die Scheidungsraten” in the desc.
- In “Der Spiegel”: Literaturnobelpreistrager, Literatur-Nobelpreistrager, Literaturnobelpreis-Tragerin; Literaturnobelpreis v.s. “Nobelpreis fur Literatur”.
- The German translation of “Latin America” by Babelfish was “lateinischem Amerika”, not “Lateinamerika”.
- “Bronchialasthma” was not translated into English by Babelfish, but “Bronchial” and “Asthma” were.

German/Dutch Decompounding Procedure

- Create a German/Dutch base dictionary consisting of single words only (compounds are excluded).
- Decompose a compound into component words found in the German/Dutch base dictionary.
- Choose the decomposition with the minimum number of component words.
- If there are more than one decompositions having the minimum number of component words, choose the decomposition with the highest probability.

German Decompounding: Example 1

Compound: fussballeuropameisterschaft (European Football Cup)

1. Base dictionary

…
ball
europa
fuss
fussball
meisterschaft
s
…

2. Decompose a compound with respect to the base dictionary.

1. fuss ball europa meisterschaft
2. fussball europa meisterschaft

3. Choose the decomposition with the smallest number of component words.

fussballeuropameisterschaft =
fussball europa meisterschaft

German Decompounding: Example 2

Compound: wintersports (winter sports)

1. Base dictionary

…
port
ports
s
sport
sports
winter
winters
…

2. Decompose a compound with respect to the base dictionary.

Decompositions	log p(D)
1. winter s ports	-43.7
2. winter sports	-20.1
3. winters ports	-28.4

3. Choose the most likely decomposition.

wintersports = winter sports

Decompounding: Probability of Decomposition

$$C = W_1 \; W_2 \; W_3 \; W_4$$

$$p(C) = p(W_1) * p(W_2) * p(W_3) * p(W_4)$$

$$p(w) = \frac{tfc(w)}{\sum_{i=1}^{n} tfc(w_i)}$$

Relative frequency of w in a collection.

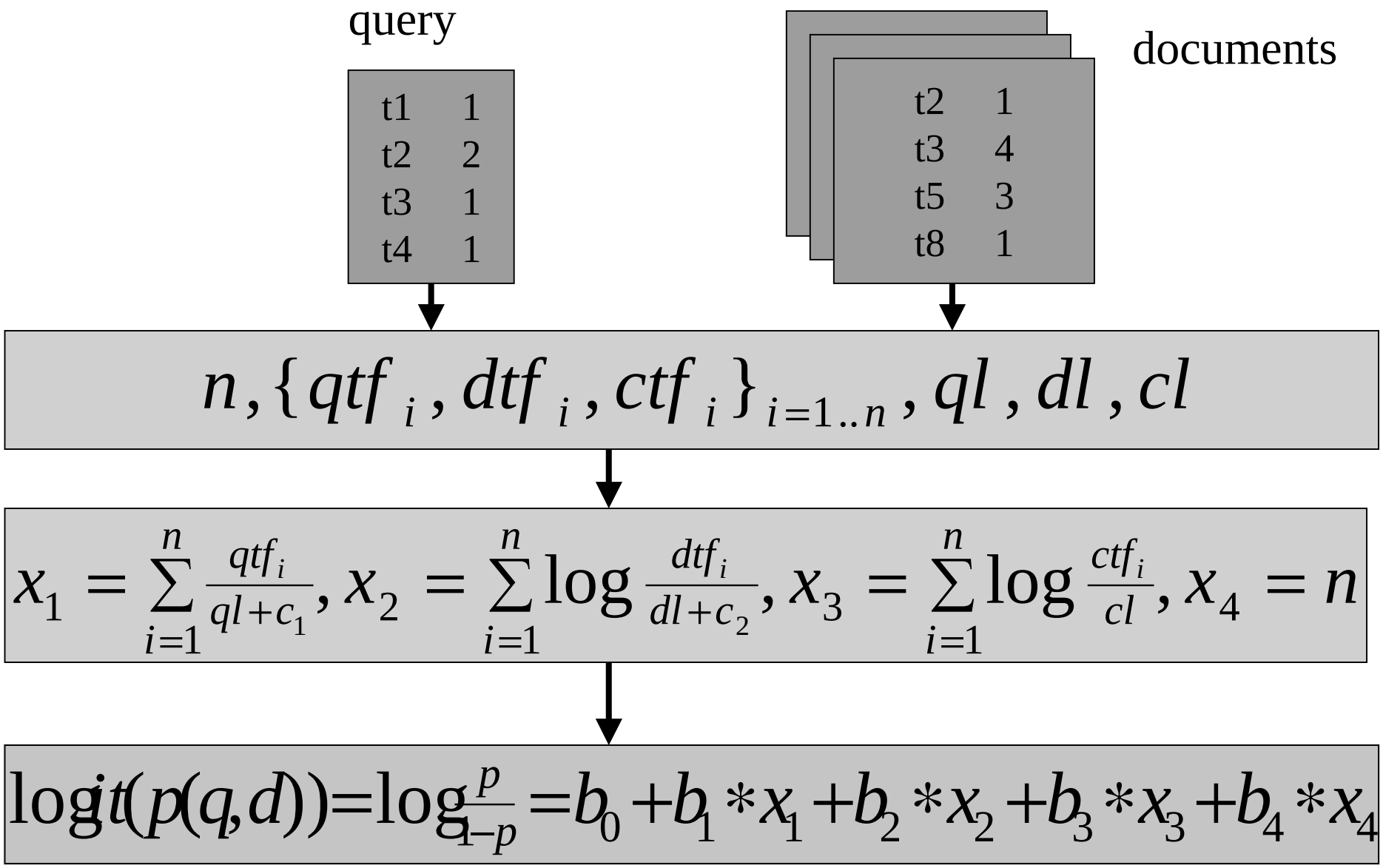
tfc(w) is the number of times word w occurs in a corpus.

n is the number of unique words (including compounds) in a corpus.

Evaluation of Decompounding

Test collection	Run type	Without decompounding	With decompounding	Change
CLEF-2002	German--German	0.3462	0.3859	+11.47%
CLEF-2002	Dutch--Dutch	0.4021	0.4186	+04.10%
CLEF-2001	Dutch--Dutch	0.3239	0.3676	+13.49%
CLEF-2002	English--German (L&H)	0.2776	0.3009	+08.4%
CLEF-2002	English--German (Babelfish)	0.2554	0.2906	+13.87%
CLEF-2002	French—German (Babelfish)	0.2774	0.3092	+11.46%

Document Ranking



Query Expansion

- Select terms from top-ranked documents after the initial search.
- Assign weight to selected terms.
- Combine selected terms with original query terms.

Query Expansion (2)

Step 1: term selection.

$$w = \frac{n_1 / n_2}{n_3 / n_4}$$

	relevant	irrelevant
indexed	n1	n3
not indexed	n2	n4

1) Rank terms in the presumed relevant documents by w in descending order; 2) Choose the top-ranked m terms, m = 2 * average-number-of-unique-query-terms. Alternative weighting schemes include a) Maximum Likelihood Ratio; b) Chi-square statistics; c) Mutual information.

Steps 2 & 3: term-weighting and merging.

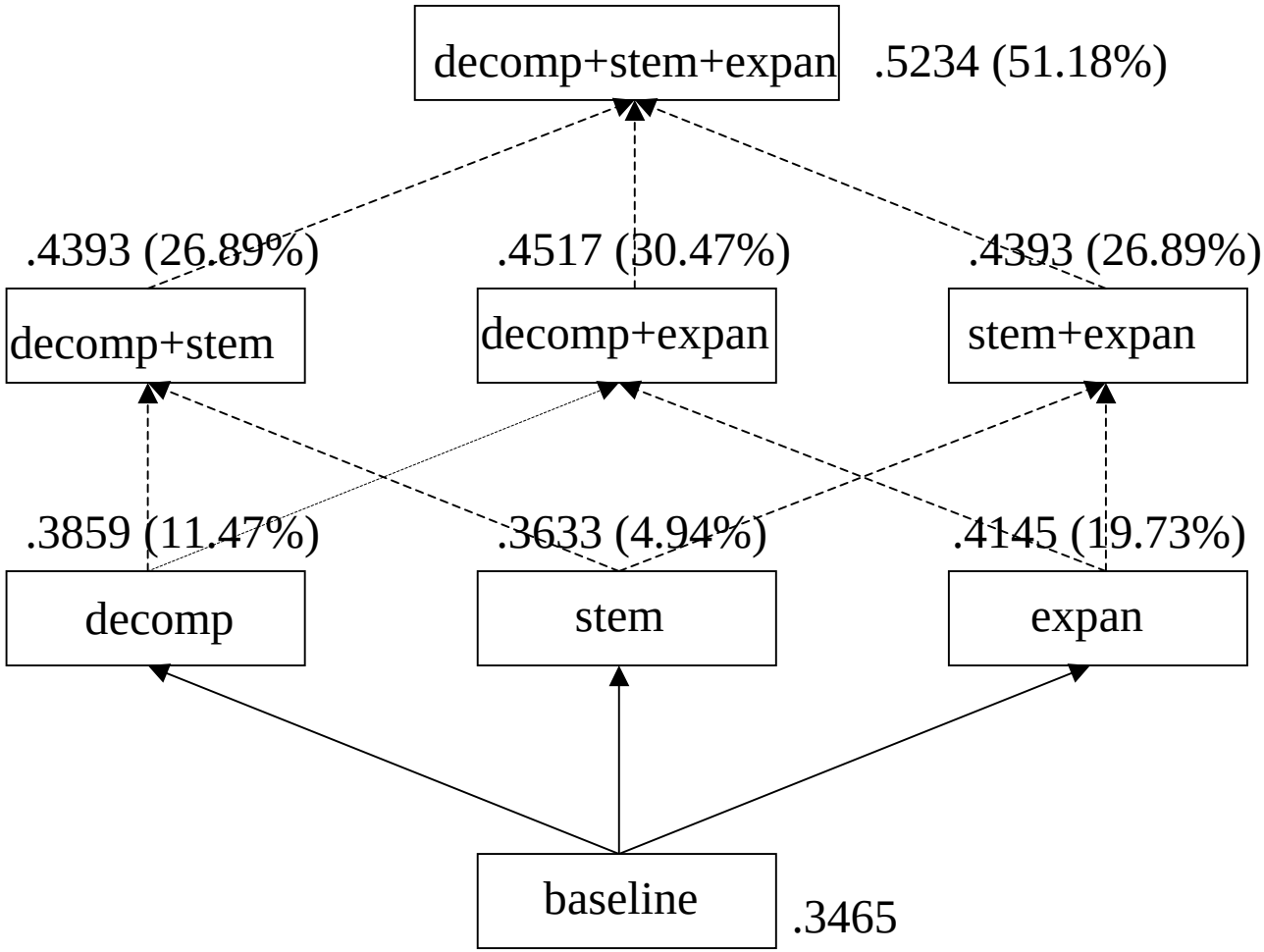
Initial query	Selected terms	Expanded query
T1 (1)		T1 (1.0)
T2 (2)	T2 (2*0.5)	T2 (3.0)
T3 (1)	T3 (1*0.5)	T3 (1.5)
	T4 (0.5)	T4 (0.5)

Evaluation of Query Expansion (10 terms/10 docs)

Run id	Run type	No query expansion	Query expansion	Change
bky2monl	Dutch-Dutch	0.4446	0.4847	+09.02%
bky2mofr	French-French	0.4347	0.5191	+19.42%
bky2mode	German-German	0.4393	0.5234	+19.14%
bky2moit	Italian-Italian	0.4169	0.4750	+13.94%
bky2moes	Spanish-Spanish	0.5016	0.5338	+06.42%
bky2bienfr	English-French	0.4118	0.4773	+15.91%
bky2bienfr2	English-French	0.4223	0.4744	+12.34%
bky2bidefr	German-French	0.3437	0.4124	+19.99%
bky2biende	English-German	0.3561	0.4479	+25.78%
bky2bifrde	French-German	0.3679	0.4759	+29.36%
bky2bienit	English-Italian	0.3608	0.4090	+13.36%
bky2bienes	English-Spanish	0.4090	0.4567	+11.66%
bky2biennl	English-Dutch	0.2564	0.3199	+24.77%

Evaluation of Decompounding, Stemming and Query Expansion in Monolingual Retrieval

(Topics: German, TD)



English-to-French Dictionary Built from Parallel Texts

1. fall in sale of cars

automne	0.32
tomber	0.08
telever	0.06

2. ski race; car race

race	0.60
courser	0.18
racial	0.05

3. pop star; galaxy star

star	0.62
etoile	0.13
etoiler	0.08

4. rock music

rock	0.89
rocher	0.02
pierre	0.01

5. lead singer

mener	0.13
conduire	0.08
amener	0.07
…	
principal	

Run id	Resource	AP
bky2enfr3	Babelfish	0.4583
bky2enfr4	L&H	0.4652
bky2enfr5	Parallel texts	0.4529

Conclusions

- The simplest direct merging method worked better than the score-normalized method when the intermediate ranked lists were produced under similar conditions (e.g., roughly the same query length, the same number of terms selected from the same number of documents).
- Decompounding improved the retrieval performance of German/Dutch monolingual and cross-language retrieval to German. The margin of improvement varies from one topic set to another.
- Query expansion substantially improved the performance of monolingual, cross-language, and multilingual retrieval.

THANK YOU