Neuchatel at NTCIR-4
From CLEF to NTCIR

Jacques Savoy
University of Neuchatel, Switzerland
www.unine.ch/info/clef/

From CLEF to NTCIR

European languages, Asian languages,
different languages
but same IR problems?

one byte = one char
limited set of char
space between words
different writings

But
same indexing?
same search and
translation scheme?

Indexing methods

- E: Words
  - Stopword list
  - Stemming

SMART system

- CJK: bigrams
  - Stoplist
  - No stemming

In K, 80% of nouns are composed of two characters (Lee *et al.*, IP&M, 1999)

Example in Chinese

我不是中国人

我不 不是 是中

中国 国人

IR models

- Probabilistic
  - Okapi
  - Prosit or deviation from randomness
- Vector-space
  - Lnu-ltc
  - tf-idf (ntc-ntc)
  - binary (bnn-bnn)

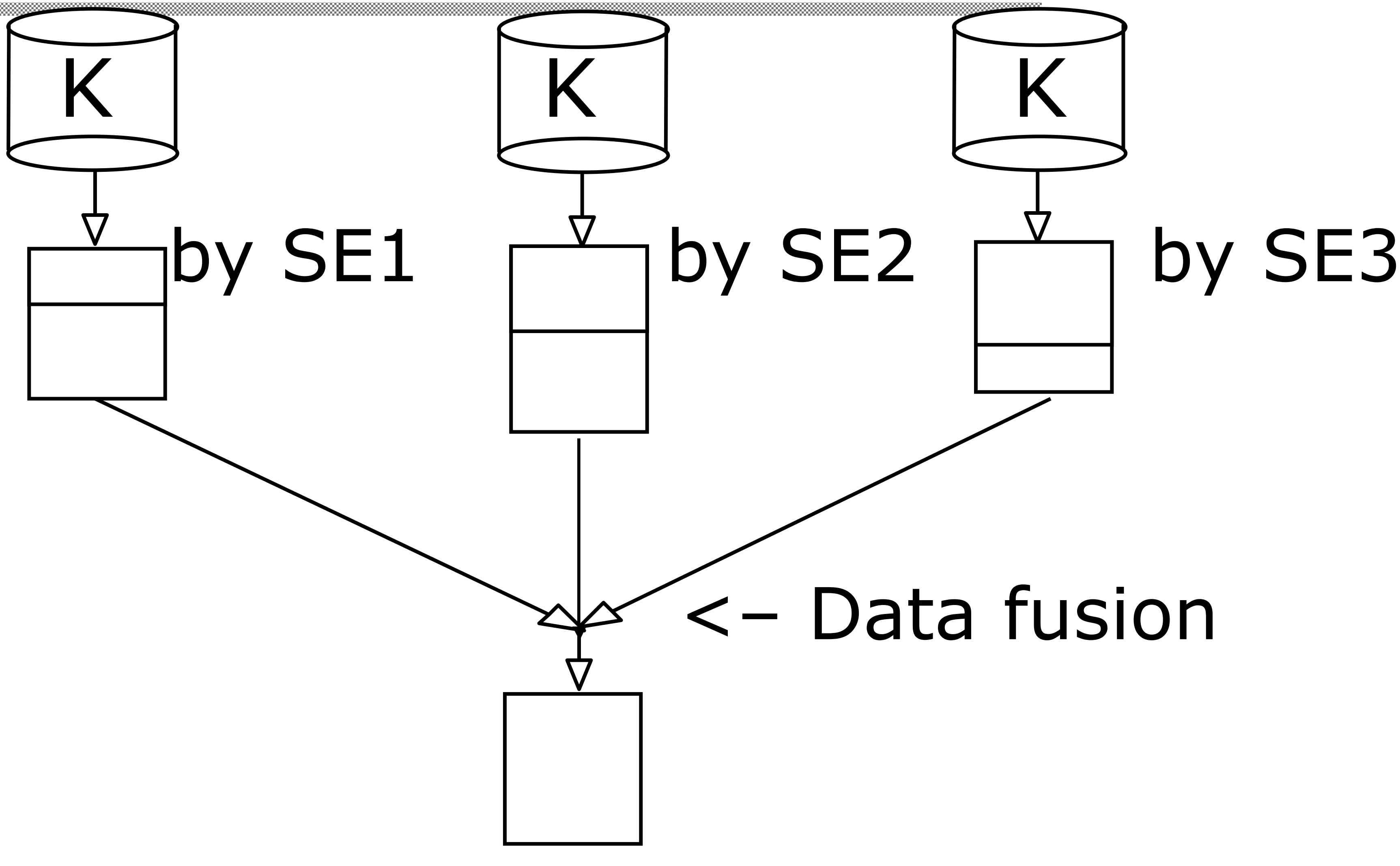
Monolingual evaluation

Model	English		Korean	
	T	D	T	D
Okapi	**0.3132**	0.2992	0.4033	0.3475
Prosit	0.2997	0.2871	0.3882	0.3010
Lnu-ltc	0.3069	**0.3139**	**0.4193**	**0.4001**
tf-idf	0.1975	0.2171	0.3245	0.3406
binary	0.1562	0.1262	0.1944	0.0725

Monolingual evaluation

Model	English		Korean	
	T	D	T	D
Okapi	0.3132	0.2992	0.4033	0.3475
+PRF	0.3594	0.3181	**0.4960**	**0.4441**
	+15%	+6%	+23%	+28%
Prosit	0.2997	0.2871	0.3882	0.3010
+PRF	**0.3731**	**0.3513**	0.4875	0.4257
	+25%	+22%	+26%	+41%

Data Fusion



Data fusion

1	KR120	1.2
2	KR200	1.0
3	KR050	0.7
4	KR705	0.6
…		

1	KR043	0.8
2	KR120	0.75
3	KR055	0.65
4	…	

1	KR050	1.6
2	KR005	1.3
3	KR120	0.9
4	…	

1 KR…
2 KR…
3 KR…
4 ….

Data fusion

- Round-robin (baseline)
- Sum RSV (Fox et al., TREC-2)
- Normalize (divide by the max)
- Z-score

Z-score normalization

1	KR120	1.2
2	KR200	1.0
3	KR050	0.7
4	KR765	0.6
…		…

Compute the mean μ and standard deviation σ

New score =
((old score-μ) / σ) + δ

1	KR120	7.0
2	KR200	5.0
3	KR050	2.0
4	KR765	1.0
…		

Monolingual (data fusion)

Korean best single	T (4 SE) 0.4868	TDNC (2 SE) **0.5141**
Round-robin	0.4737	0.5047
SumRSV	0.5044	0.5030
Norm max	**<u>0.5084</u>**	0.5045
Z-score	<u>0.5074</u>	0.5023
Z-score wt	<u>0.5078</u>	0.5058

Monolingual evaluation (C)

Model	Chinese-unigram		Chinese-bigram	
	T	D	T	D
Okapi	0.1667	0.1198	0.1755	0.1576
Prosit	0.1452	0.0850	0.1658	0.1467
Lnu-ltc	**0.1834**	**0.1484**	**0.1794**	**0.1609**
tf-idf	0.1186	0.1136	0.1542	0.1507
binary	0.0431	0.0112	0.0796	0.0686

Monolingual evaluation (C)

Model	Chinese-unigram		Chinese-bigram	
	T	D	T	D
Okapi +PRF	0.1667 **0.1884** +13%	0.1198 **0.1407** +17%	0.1755 0.2004 +14%	0.1576 0.1805 +15%
Prosit +PRF	0.1452 0.1659 +14%	0.0850 0.1132 +33%	0.1658 **0.2140** +29%	0.1467 **0.1987** +35%

Monolingual evaluation (J)

Model	Bigram (kanji,kata)		Bigram (kanji)	
	T	D	T	D
Okapi	**0.2873**	**0.2821**	**0.2972**	**0.2762**
Prosit	0.2637	0.2573	0.2734	0.2517
Lnu-ltc	0.2701	0.2740	0.2806	0.2718
tf-idf	0.2104	0.2087	0.2166	0.2101
binary	0.1743	0.1741	0.1703	0.1105

Monolingual evaluation (J)

Model	Bigram (kanji,kata)		Bigram (kanji)	
	T	D	T	D
Okapi	0.2873	0.2821	0.2972	0.2762
+PRF	0.3259	0.3331	**0.3514**	0.3200
	+13%	+18%	+18%	+16%
Prosit	0.2637	0.2573	0.2734	0.2517
+PRF	**0.3396**	**0.3394**	0.3495	**0.3218**
	+29%	+32%	+28%	+28%

Translation resources

- Machine-readable dictionaries
  - Babylon
  - Evdict
- Machine translation services
  - WorldLingo
  - BabelFish
- Parallel and/or comparable corpora (not used in this evaluation campaign)

Bilingual evaluation E->C/J/K

T Okapi	Chinese bigram	Japanese bigram k&k	Korean bigram
Manual	0.1755	0.2873	0.4033
Babylon 1	0.0458	0.0946	0.1015
Lingo	0.0794	0.1951	0.1847
Babelfish	0.0360	0.1952	**0.1855**
Combined	**0.0854**	**0.2174**	0.1848

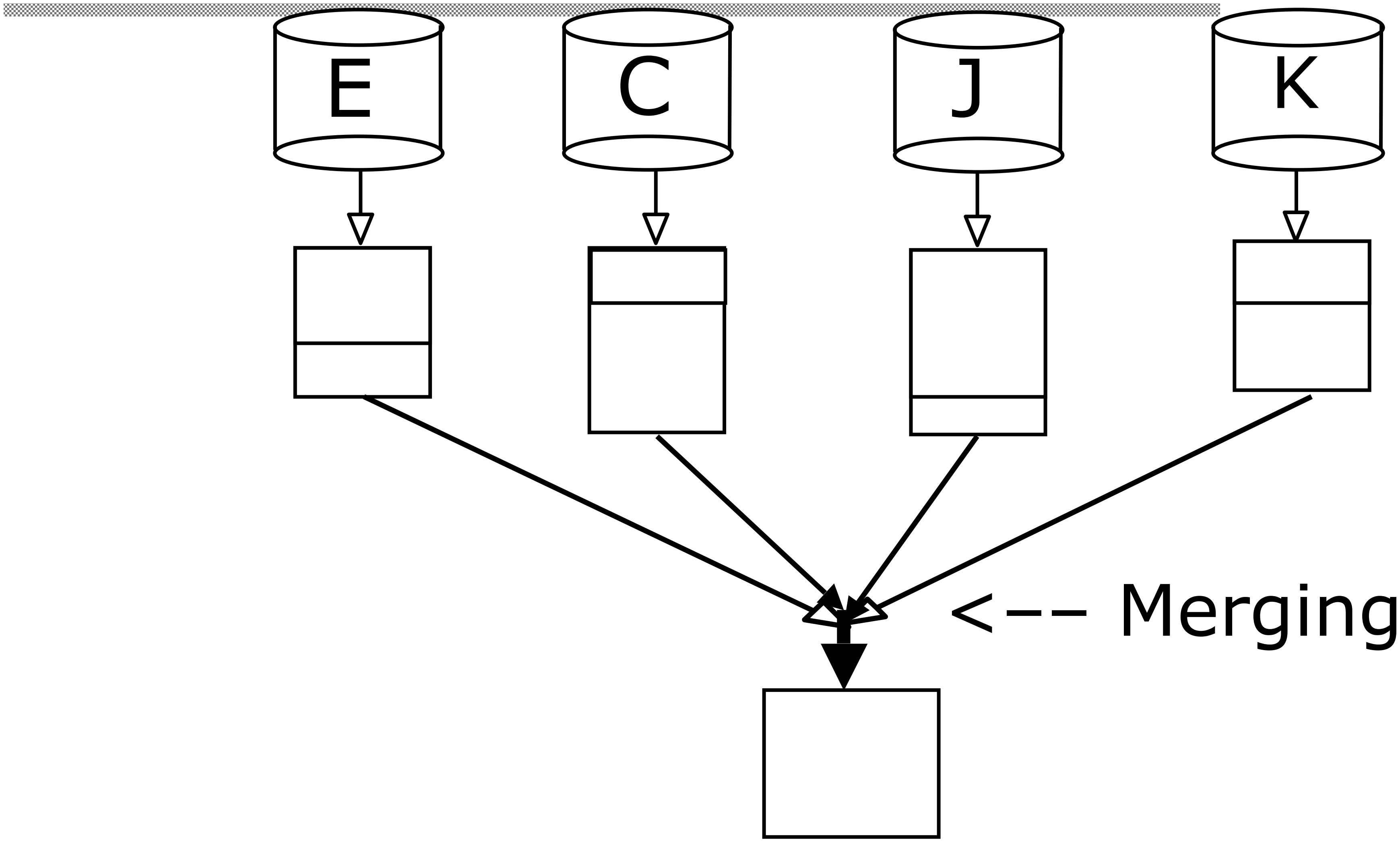
Bilingual evaluation E->C/J/K

T	Chinese bigram	Japanese bigram k&k	Korean bigram
Manual	0.1755	0.2873	0.4033
Okapi	0.0854	0.2174	0.1848
+PRF	0.1039	**0.2733**	**0.2397**
Prosit	0.0817	0.1973	0.1721
+PRF	**0.1213**	0.2556	0.2326

Multilingual IR E->CJKE

- Create a common index
  Document translation (DT)
- Search on each language and
  merge the result lists (QT)
- Mix QT and DT
- No translation

Merging problem



Multilingual IR (merging)

- Round-robin (baseline)
- Raw-score merging
- Normalize (by the max)
- Z-score
- Logistic regression

Test-collection NTCIR-4

	E	C	J	K
size	619 MB	490 MB	**733** MB	**370** MB
doc	347550	381681	596058	254438
mean	96.6	363.4	114.5	236.2
topic	58	59	55	57
rel.	35.5	19	**88**	**43**

Multilingual evaluation

CJE	T (auto)	T (manual)
Round-robin	0.1564	0.2204
Raw-score	0.1307	0.2035
Norm max	0.1654	0.2222
Biased RR	0.1413	0.2290
Z-score wt	**0.1719**	**0.2370**

Multilingual evaluation

CJKE	T (auto)	T (manual)
Round-robin	0.1419	0.2371
Raw-score	<u>0.1033</u>	<u>0.1564</u>
Norm max	0.1411	0.2269
Biased RR	<u>0.1320</u>	0.2431
Z-score	<u>**0.1446**</u>	<u>**0.2483**</u>

Conclusions (monolingual) From CLEF to NTCIR

- The best IR model seems to be language-dependant (Okapi in CLEF)
- Pseudo-relevance feedback improves the initial search
- Data fusion (yes, with shot queries
  limited in CLEF)

Conclusions (bilingual) From CLEF to NTCIR

- Translation resources freely available produce a poor IR performance (differs from CLEF)
- Improvement by
  - Combining translations (not here, yes in CLEF)
  - Pseudo-relevance feedback (as in CLEF)
  - Data fusion (not clear)

Conclusions (multilingual) From CLEF to NTCIR

- Selection and merging are still hard problems (as in CLEF)

- Z-score seems to produce good IR performance over different conditions (as in CLEF)