

Word or bigram for effective search with Asian languages ?

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

Overview

- Which IR model performs the best?
- Are bigrams more effective than words?
- Pseudo-relevance feedback & failure
- Data fusion
- Bilingual and multilingual searches

Which IR model ?

Comparisons made using the English language corpus (rigid evaluation)

Using eleven search models:

2 probabilistic model

9 vector-space models

Analysis of differences obtained using a statistical test (5%, two-sided)

The best IR model

| English | T | D | DN |
|---------------|---------------|---------------|---------------|
| Okapi | 0.3692 | 0.3615 | 0.4555 |
| $l(n)L2$ | <u>0.3591</u> | <u>0.3548</u> | 0.4556 |
| Lnu-ltc | 0.3562 | 0.3551 | <u>0.4185</u> |
| dtu-dtn | <u>0.3577</u> | 0.3748 | <u>0.3949</u> |
| <i>tf idf</i> | <u>0.2345</u> | <u>0.2522</u> | <u>0.3061</u> |

Which IR model ?

Okapi or *Deviation from randomness*

Relative improvement (prob. vs vector-space)

+21% with T queries

+14% with D queries

Okapi vs. *tf idf* model

+53% with T queries

+41% with D queries

Word or bigram?

In the Japanese or Chinese language, word boundaries are not explicitly marked.

Each sentence is automatically segmented using the *Chasen* (JA) or *Mandarin Tools* (ZH)

In Korean, words are decompounded using the *Hangul Analysis Module*

Word or bigram?

我不是中国人

Correct segmentation

我 不 是 中 国 人

Word or bigram?

我不是中国人

Bigrams generation

我 不 不 是 是 中
中 国 国 人

Word or bigram (JA)?

| Japanese | bigram T | word T | bigram DN | word DN |
|---------------|---------------|---------------|---------------|---------------|
| Okapi | 0.2660 | <u>0.2655</u> | 0.4079 | 0.4002 |
| PB2 | 0.2717 | 0.2895 | <u>0.3957</u> | 0.3925 |
| <i>tf idf</i> | <u>0.1292</u> | <u>0.1227</u> | <u>0.2302</u> | <u>0.1987</u> |

Word or bigram (ZH)?

| Chinese | bigram T | word T | bigram DN | word DN |
|---------------|---------------|---------------|---------------|---------------|
| Okapi | 0.2995 | 0.3230 | 0.3887 | 0.4135 |
| PB2 | 0.3042 | 0.3246 | 0.3973 | 0.4136 |
| <i>tf idf</i> | <u>0.2130</u> | <u>0.1645</u> | <u>0.3138</u> | <u>0.2741</u> |

Word or bigram (KR)?

| Korean | bigram T base | word T | word + decomp T |
|--------|---------------------|---------------|-----------------------|
| Okapi | 0.3630 | <u>0.2245</u> | 0.3549 |
| PB2 | 0.3729 | <u>0.2378</u> | 0.3659 |

Processing time / space

Bigram indexing is more complex

of postings 11x for Chinese
 3x for Japanese

Inverted file size 2x for Chinese
 1.5x for Japanese

Building time 2x for Chinese
 1.3 for Japanese

Processing time / space

When searching

N-gram querying can be 10 times slower [McNamee, CLEF 2005]

Similar for both Chinese & Japanese

Blind query expansion

Adding terms from the top-ranked documents usually improves MAP

But this method failed for Genomics TREC 2005!

We used

Rocchio's approach

our idf-based strategy

$$w_j = \alpha \text{tf}_{ij} + (\alpha/k) \text{idf}_j$$

Blind query expansion

| T queries | English word | Japane. word | Chinese word | Korean bigram |
|-----------|---------------|---------------|---------------|---------------|
| Okapi | 0.3692 | 0.2655 | 0.3230 | 0.3630 |
| Rocchio | 0.4420 | 0.3523 | 0.3788 | 0.4346 |
| IDFQE | 0.4476 | 0.3681 | 0.3778 | 0.4453 |

But sometimes we fail!

Topic #45

<title> **population issue, hunger**

<desc> Find documents describing the **effects population issues** have on **hunger**.

But sometimes we fail!

Five relevant articles for Query #45

<HEADLINE> Readers Forum: What do you think about genetically modified (GM) food?

<HEADLINE> Editorial / Serious food for thought

<HEADLINE> Bangladeshi Population May Reach 210.8 Million in 50 Ye.

<HEADLINE> News Analysis: Major Trends in World Population Growth (1) by Gu Zhenqiu

<HEADLINE> 17 Sub-Saharan African Countries Facing Food Emergencies: FAO

But sometimes we fail!

| English Query #45 | T | D | DN |
|-------------------|--------|--------|--------|
| Okapi | 0.0020 | 0.0017 | 0.0274 |
| $l(n)L2$ | 0.0008 | 0.0007 | 0.0237 |
| PB2 | 0.0031 | 0.0031 | 0.0531 |
| <i>tf idf</i> | 0.0001 | 0.0001 | 0.0011 |

Why we fail?

Query #45

<title> **population issue, hunger**

<desc> Find documents describing the effects **population issues** have on **hunger**.

populat -> df high (7,995)

issu -> df very high (44,209)

hung -> (df=3,036) but it does not appear in relevant items

Why we fail?

The context may be useful.

Query phrase may help ... "population issue"

The query phrase "population issue" does not appear in the relevant documents.

Why we fail?

In the relevant articles, we may find

"food is expected to be a major issue in the next century"

"political issue", "food issues"

"issue of food security", "as a strategic issue"

"remain several unresolved issues."

Why we fail? (2nd example)

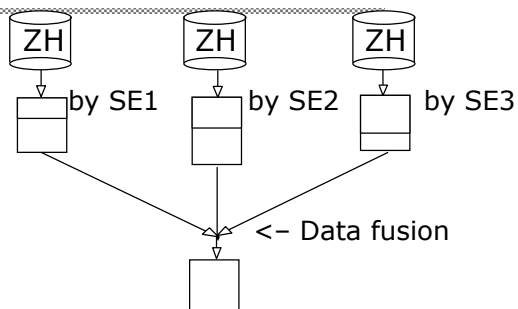
Query #4

<title> the US Secretary of Defense, William Sebastian Cohen, Beijing

And the top-ranked documents, all non-relevant

1. <HEADLINE> U.S. Defense Secretary Arrives in China
2. <HEADLINE> Weekly Highlights of Diplomatic News
3. <HEADLINE> U.S. Secretary of Defence to Visit China
4. <HEADLINE> U.S. Defense Secretary Arrives in China
5. <HEADLINE> Highlights of Diplomatic News Today

Data Fusion



Data fusion

- Round-robin (baseline)
- Sum RSV (Fox et al., TREC-2)
- Normalize (divide by the max)
- Z-score
- Effective? In some cases, yes
e.g., NLM at Genomics TREC-2005

Monolingual (data fusion)

| | ZH T (3 SE) | JA T (2 SE) |
|-------------|---------------|---------------|
| best single | 0.3912 | 0.3681 |
| Round-robin | 0.3780 | 0.3639 |
| SumRSV | 0.4121 | 0.3637 |
| Norm max | 0.4062 | 0.3734 |
| Z-score wt | 0.4050 | 0.3754 |

Translation resources

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

Bilingual evaluation E->C/J/K

| T | Chinese bigram | Japanese bigram | Korean bigram |
|--------------------------|----------------|-----------------|---------------|
| Manual | 0.2995 | 0.2660 | 0.3630 |
| alphaWorks | <u>0.1208</u> | <u>0.1855</u> | <u>0.2055</u> |
| 3 MTs | <u>0.1317</u> | <u>0.1927</u> | <u>0.2396</u> |
| 3 MTS + PRF & Datafusion | <u>0.2417</u> | 0.2631 | <u>0.3374</u> |

Bilingual evaluation (failure)

Missing translations of proper nouns (e.g., person, geographic, product name, ...)

E.g., the three hardest topics (E->J)

Q#31 Fine dust particles, heart disease

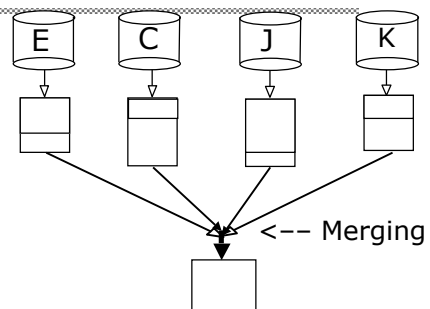
Q#35 Capital punishment, survey data

Q#28 Bubka, human bird, retirement

Multilingual IR

- Search on each language and merge the result lists (QT)

Merging problem



Multilingual IR

- Search on each language and merge the result lists (QT)
- Round-robin (baseline)
- Raw-score merging
- Normalize (by the max)
- Z-score

Multilingual evaluation

| E -> CJKE | T (data fusion) | T (Okapi) |
|-------------|-----------------|---------------|
| Round-robin | 0.2244 | 0.2169 |
| Raw-score | 0.2165 | 0.2332 |
| Norm max | 0.2248 | 0.2102 |
| Biased RR | <u>0.2036</u> | <u>0.1965</u> |
| Z-score | 0.2333 | <u>0.2261</u> |

Test-collection NTCIR-5

| | E | C | J | K |
|-----------------|----------------|----------------|----------|---------|
| size | 438 MB | 1,100 MB | 1,100 MB | 312 MB |
| # doc. | 259,050 | 901,446 | 858,400 | 220,374 |
| # topic | 49 | 50 | 47 | 50 |
| mean rel. items | 62.7 | 37.7 | 44.9 | 36.6 |

Conclusions (monolingual)

- The best IR model seems to be language-independent (Okapi, DFR)
- Pseudo-relevance feedback improves (IDFQE seems better than Rocchio - to be confirmed)
- Data fusion (not sure!)
- There is always hard topics

Conclusions (monolingual)

- Words or bigrams seem to present the same performance for JA & ZH (words slightly better for ZH)
- Bigrams seem the best approach for KR (to verify!)
- Processing space / time