NTCIR-5
Evaluation Campaign

University of Neuchâtel

TEAM Information Retrieval

Monolingual in English, Japanese, Chinese and Korean languages

Main findings

- Okapi (or DFR) are the best performing models for the Japanese and Chinese languages (bigram or word-based). Relative improvement over the classical *tf idf* : +32% Korean, +35% for Chinese, +48% English, and +111% Japanese language.
- Lnu or dtu propose also good performances for the English and Korean languages.
- Blind query expansion statistically improves the MAP (for T queries, +15% for Korean, +19% for Chinese, +22% for English, and +28% for Japanese language).
- Our idf-based query expansion (IDFQE) produces usually a better performance level than Rocchio's approach (but the differences are usually not significant!).
- Data fusion may improve the MAP (the performance differences are not statistically significant when compared to the best single system).

Monolingual evaluation

Model	English	Japanese	Chinese	Korean
T query	word	word	word	bigram
I(*n*)L2/PB2	0.3591	**0.2895**	**0.3246**	0.3729
Okapi-npn	**0.3692**	0.2655	0.3230	0.3630
Lnu-ltc	0.3562	0.2743	0.3227	**0.3973**
dtu-dtn	0.3577	0.2735	0.2894	0.3673
atn-ntc	0.3423	0.2109	0.2578	0.3270
ltn-ntc	0.3275	0.2723	0.2833	0.3708
ntc-ntc	0.2345	0.1227	0.1645	0.2506
ltc-ltc	0.2509	0.0945	0.1772	0.2260
lnc-ltc	0.2617	0.1132	0.2189	0.2414
bnn-bnn	0.2000	0.1403	0.1542	0.2348
nnn-nnn	0.1048	0.1055	0.0738	0.1770

Statistically significant improvement over the baseline (in bold) are underlined.

Blind query expansion

Model	English	Japanese	Chinese	Korean
T query	word	word	word	bigram
I(*n*)L2/PB2	0.3591	0.2895	0.3246	0.3729
#doc/#term	15 / 100	15 / 100	5 / 75	15 / 140
& Rocchio	0.4450	0.3479	0.3547	0.3899
#doc/#term	15 / 100	15 / 100	5 / 125	15 / 100
& idfqe	0.4389	**0.3690**	0.3769	0.4253
Okapi-npn	0.3692	0.2655	0.3230	0.3630
#doc/#term	15 / 100	15 / 100	5 / 75	15 / 100
& Rocchio	0.4420	0.3523	**0.3788**	0.4346
#doc/#term	15 / 100	20 / 100	5 / 125	15 / 100
& idfqe	**0.4476**	0.3681	0.3778	**0.4453**

Statistically significant improvement over the baseline (without query expansion) are underlined.

Some weighting schemes

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$	atn	$w_{ij} = idf_j \cdot [0.5 + 0.5 \cdot tf_{ij} / \max tf_{i.}]$
dtn	$w_{ij} = [\ln(\ln(tf_{ij}) + 1) + 1] \cdot idf_j$	npn	$w_{ij} = tf_{ij} \cdot \ln[(n - df_j) / df_j]$
Okapi	$w_{ij} = \left((k_1 + 1) \cdot tf_{ij}\right) \Big/ \left(K + tf_{ij}\right)$	Lnu	$w_{ij} = \frac{\left(1 + \ln(tf_{ij}) \big/ \ln(\text{mean } tf) + 1\right)}{(1 - slope) \cdot pivot + slope \cdot nt_i}$
lnc	$w_{ij} = \frac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} \left(\ln(tf_{ik}) + 1\right)^2}}$	ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(tf_{ik} \cdot idf_k\right)^2}}$
ltc	$w_{ij} = \left(\ln(tf_{ij}) + 1\right) \cdot idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \cdot idf_k\right)^2}$	dtu	$w_{ij} = \frac{\left(\ln\left(\ln(tf_{ij}) + 1\right) + 1\right) \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$

Example of hard topic

<NUM> 045
<TITLE> population issue, hunger
<DESC> Find documents describing the effects population issues have on hunger.

As query: [issue (df=44,209), population (df=7,995), hunger (df=3,036)]

The Okapi model retrieves two relevant items (over 5) at position 478 and 547.

Example of document

```
<DOCNO> XIE20000806.0034 </DOCNO>
<LANG> EN </LANG>
<HEADLINE> Bangladeshi Population May Reach 210.8 Million in 50 Years </HEADLINE>
<DATE> 2000-08-06 </DATE>
<TEXT>
When the world's population hits 9 billion by the middle of the new century at the current growth rate, Bangladesh is likely to be crammed with 210.8 million people after 50 years, according to a data sheet of the U.S.-based Population Reference Bureau (PRB) received here Sunday.
Ranking eighth at present, Bangladesh will remain pegged to the same status while some of the countries would have shifted from their present positions.
The just-released World Population Data Sheet 2000 containing the demographic data worldwide showed that the world is expected to add 3 billion more people to reach a total of 9 billion by 2050.
And the PRB, in its press release, made the grim predictions that the current period of rapid population growth would continue for at least another 50 years to 2050.
…</TEXT>
```

Word or bigram for effective Japanese, Chinese or Korean retrieval

Main findings

- For the Japanese language
  - bigram and word have a similar performance level (from a statistical point of view).
  - Small improvement when using words, e.g., 4% with T queries.
- For the Chinese language
  - bigram and word usually proposes statistically the same performance level (e.g., relative difference around 1% (T queries) or 3% (DN queries)).
- For the Korean language
  - bigram is better than word (differences are statistically significant). For T queries, the relative difference favors bigram-based scheme by around 37%.
  - decompounding is better than word (differences are statistically significant, e.g., +56% for T queries, or +70% for DN queries).
  - bigram is better than decompounding (differences not always significant). The relative difference favors bigram-based scheme by around 2% (T queries) or 9% (D queries)

Mean average precision (Japanese corpus)

Model	T	T	D	D	DN	DN
	bigram	word	bigram	word	bigram	word
PB2	0.2717	**0.2895**	0.2829	**0.3120**	0.3957	0.3925
Okapi base	0.2660	0.2655	0.2694	0.2657	**0.4079**	0.4002
PB2	0.2717	0.2895	0.2829	0.3120	0.3957	0.3925
& Rocchio	0.3429	0.3479	0.3596	0.3581	0.4240	0.3983
& IDFQE	0.3476	**0.3690**	0.3563	0.3609	0.4180	0.4071
Okapi	0.2660	0.2655	0.2694	0.2657	0.4079	0.4002
& Rocchio	0.3266	0.3523	0.3212	0.3433	0.4103	0.4021
& IDFQE	0.3501	0.3681	0.3617	**0.3763**	0.4307	**0.4378**

Mean average precision (Chinese corpus)

Model	T	T	D	D	DN	DN
	bigram	word	bigram	word	bigram	word
PB2	0.3042	**0.3246**	0.2878	**0.2974**	0.3973	**0.4136**
Okapi	0.2995	0.3230	0.2584	0.2816	0.3887	0.4135
PB2	0.3042	0.3246	0.2878	0.2974	0.3973	0.4136
& Rocchio	0.3782	0.3547	0.3616	0.3822	0.4241	0.4088
& IDFQE	**0.3912**	0.3769	0.3861	**0.3954**	0.4288	0.4400
Okapi	0.2995	0.3230	0.2584	0.2816	0.3887	0.4135
& Rocchio.	0.3559	0.3788	0.3176	0.3522	0.3854	0.4252
& IDFQE	0.3557	0.3778	0.3659	0.3576	0.4242	**0.4479**

Mean average precision (Korean corpus)

Model	T	T	T	D	D	D
	bigram	word	HAM	bigram	word	HAM
PB2	**0.3729**	0.2378	0.3659	**0.4141**	0.1824	0.3786
Okapi	0.3630	0.2245	0.3549	0.3823	0.1716	0.3447

The baseline is the bigram performance. Differences that are statistically significant are underlined.
We have used the morphological analyzer ChaSen for the Japanese, the Mandarin Tools (freely available at www.mandarintools.com) for the traditional Chinese, and the Hangul Analysis Module (HAM) for the Korean language.

Automatic segmentation vs. bigram

我不是中国人

The correct segmentation

我　不　是　中国人

The bigrams

我不　不是　是中

中国　国人

Contact:

Prof. Jacques Savoy, University of Neuchatel
Rue Pierre-à-Mazel 7, CH-2000 Neuchâtel
Web : http://www.unine.ch/info/clef
e-mail : Jacques.Savoy@unine.ch