Experiments in Finding Chinese and Japanese Answer Documents at NTCIR-7

Stephen Tomlinson
Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com

November 7, 2008

Abstract

We describe evaluation experiments conducted by submitting retrieval runs for the natural language Simplified Chinese, Traditional Chinese and Japanese questions of the Information Retrieval for Question Answering (IR4QA) Task of the Advanced Cross-lingual Information Access (ACLIA) Task Cluster of the 7th NII Test Collection for IR Systems Workshop (NTCIR-7). In a sampling experiment, we found that, on average per topic, the percentage of answer documents assessed was less than 65% for Simplified Chinese, 32% for Traditional Chinese and 41% for Japanese. However, our preferred measure for this task, Generalized Success@10, only considers the rank of the first answer document retrieved for each topic, as one good document answering the question is all that a user needs for this task. We experimented with different techniques (words vs. n-grams, removing question words and blind feedback) and found that the choice of technique can have a substantial impact on the rank of the first answer document for particular questions. **Keywords:** *Simplified Chinese, Traditional Chinese, Japanese, evaluation, robust retrieval, extreme topics, sampling.*

1 Introduction

Livelink ECM - eDOCS SearchServer$^{\text{TM}}$ is a toolkit for developing enterprise search and retrieval applications. The SearchServer kernel is also embedded in other components of the Livelink ECM - eDOCS Suite[1].

SearchServer works in Unicode internally [3] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (NTCIR [4], CLEF [2] and TREC [7]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

This paper describes experimental work with SearchServer for the task of finding documents containing answers to natural language questions posed in Chinese and Japanese. The test collections are from the Information Retrieval for Question Answering (IR4QA) Task of the Advanced Cross-lingual Information Access (ACLIA) Task Cluster of the 7th NII Test Collection for IR Systems Workshop (NTCIR-7). We just study monolingual document retrieval in this work (i.e. finding answer documents in the same language as the original question).

2 Methodology

2.1 Data

The document sets of the NTCIR-7 ACLIA IR4QA Task consisted of news articles in Simplified Chinese, Traditional Chinese and Japanese. Table 1 gives their sizes. For more details, please see the NTCIR-7 ACLIA IR4QA Task overview paper [6].

Table 1. Sizes of NTCIR-7 ACLIA IR4QA Document Sets

Language	Text Size	#Documents
S. Chinese	970,802,001 bytes	545,162
T. Chinese	1,448,980,305 bytes	1,150,649
Japanese	569,980,735 bytes	419,759

The test "topics" of the NTCIR-7 ACLIA IR4QA Task consisted of natural language questions (and also, for each question, a more detailed narrative, which

[1] Livelink, Open Text$^{\text{TM}}$ and SearchServer$^{\text{TM}}$ are trademarks or registered trademarks of Open Text Corporation in the United States of America, Canada, the European Union and/or other countries. This list of trademarks is not exhaustive. Other trademarks, registered trademarks, product names, company names, brands and service names mentioned herein are property of Open Text Corporation or other respective owners.

we did not use in our experiments). For each topic, the task organizers provided a set of relevance assessments (qrels). The qrels list the documents judged to be relevant (i.e. fully satisfying the information need expressed in the topic, or, stated another way, fully answering the question), partially relevant (i.e. just partially satisfying the information need expressed in the topic) and not relevant for each of the topics. The top-half of Table 2 gives the number of topics for each language which included at least one relevant document and the average number of relevant documents for each topic (along with the lowest and highest number of relevant documents of any topic). The bottom-half of Table 2 does the same when including partially relevant documents. Note that for Traditional Chinese (CT) and Japanese (JA) there are fewer topics in the top-half of the table because a few topics just had partially relevant documents; for Simplified Chinese (CS) every topic had at least one relevant document.

Table 2. Judged Topics of NTCIR-7 ACLIA IR4QA Task

Language	Topics	Rel/Topic (R)
CS	97	55 (lo 1, hi 240)
CT	94	25 (lo 1, hi 131)
JA	97	42 (lo 1, hi 357)
		Rel/Topic (P+R)
CS	97	98 (lo 7, hi 289)
CT	95	55 (lo 5, hi 169)
JA	98	87 (lo 5, hi 363)

2.2 Indexing

The experimental post-6.0 version of SearchServer used in these experiments provided both word-based and n-gram approaches to indexing.

In the word-based approach, SearchServer segmented the text into words and split compound words (decompounding). The segmenter also performed stemming for Japanese. A short stopword list was used for each language. The lexicon-based segmenters and stemmers were based on internal linguistic component 3.7.0.15.

In the n-gram approach, typically overlapping bigrams were used for most Asian text.

2.3 Searching

For all runs, SearchServer Intuitive Searching was used, i.e. the IS_ABOUT predicate of SearchSQL, which accepts unstructured text. For example, the question for topic ACLIA1-CS-T41 was "列举全球气候变暖的危害。" (List the hazards of global warming.), and the corresponding SearchSQL query was as follows:

```
SELECT RELEVANCE() AS REL, DOCNO
FROM NTC7CS
WHERE FT_TEXT IS_ABOUT
  '列举全球气候变暖的危害。'
ORDER BY REL DESC;
```

The relevance calculation included term frequency dampening [5] and inverse document frequency.

For the blind feedback runs investigated below, 3 additional IS_ABOUT queries were issued (one for each of first 3 documents retrieved by the base run). Then the 3 result lists were merged with the base result list based on the relevance scores (weight 1 for each expansion query, weight 3 for the base run). This approach is Rocchio-like with 50% weight on the original query, 50% weight on the first 3 retrieved documents (blindly assumed relevant), and 0% weight on non-relevant documents.

2.4 Evaluation Measures

This paper refers to the following retrieval measures:

Average Precision (AP): "Precision" is the percentage of retrieved documents which are relevant. For a topic, AP is the average of the precision after each relevant document is retrieved (using zero as the precision for relevant documents which are not retrieved). In this paper, AP is based on the first 1000 retrieved items. The score ranges from 0.0 (no relevant documents found) to 1.0 (all relevant documents found at the top of the list). "Mean Average Precision" (MAP) is the mean of the average precision scores over all of the topics (i.e. all topics are weighted equally).

Success@n (S@n): For a topic, Success@n is 1 if the first relevant document is found in the first n rows, 0 otherwise. This paper lists the Success@1 (S1) measure for each run.

Generalized Success@10 (GenS@10, GenS10 or GS10): For a topic, GenS10 is 1.08^{1-r} where r is the rank of the first relevant document, or zero if no relevant document is retrieved. (This measure was introduced in [8] as "First Relevant Score" (FRS).) Compared to reciprocal rank, GenS10 falls less sharply in the early ranks; e.g. GenS10 is 1.0 at rank 1, 0.93 at rank 2, 0.86 at rank 3, etc. GenS10 is considered a generalization of Success@10 because it rounds to 1 for $r \leq 10$ and to 0 for $r > 10$.

Differences in GenS10 exceeding 0.50 for a topic are particularly important. One can show that a difference in GenS10 of at least 0.50 for a topic implies that one technique retrieved an answer document in the first 10 ranks and the other did not. Furthermore, a difference in GenS10 of at least 0.50 for a topic implies

Table 3. Mean Scores of Diagnostic Runs (Full Relevance)

Run	GenS10	S1	MAP
CS-words	0.871	60/97	0.422
CS-words+qw	0.922	66/97	0.488
CS-words+qw+bf	0.919	70/97	0.502
CS-ngram	0.897	59/97	0.421
CT-words	0.825	49/94	0.366
CT-words+qw	0.850	50/94	0.385
CT-words+qw+bf	0.853	52/94	0.410
CT-ngram	0.859	53/94	0.361
JA-words	0.866	57/97	0.423
JA-words+qw	0.921	62/97	0.506
JA-words+qw+bf	0.932	64/97	0.539
JA-ngram	0.769	48/97	0.323

a difference of least 10 in the rank of the first answer document retrieved. (The proofs are left as an exercise to the reader.)

GenS10 is the best measure for "robust retrieval" that we are aware of. For sets of topics, [9] found that mean GenS10 was the most reliable of 30 investigated retrieval measures at favoring the more robust system.

While average precision is the most commonly used measure for evaluating document retrieval, we argue that GenS10 is actually the more important measure for the IR4QA task of finding documents that answer a question because the user only needs one relevant document to learn the answer to the question.

2.5 Comparision Tables

For comparison tables such as Tables 4 and 5, the columns are as follows:

"Expt" specifies the language code and label of the experiment.

"Δ" is the difference in the mean scores from this experiment for the specified measure.

"95% Conf" is an approximate 95% confidence interval for the mean difference (calculated from plus/minus twice the standard error of the mean difference). If zero is not in the interval, the result is "statistically significant" (at the 5% level).

"vs." is the number of topics on which the experimental run scored higher, lower and tied (respectively) compared to the base run. These numbers should always add to the number of topics in the experiment.

"3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets. The first difference is the largest one of any topic (based on the absolute value). The third difference is the largest difference in the other direction (so the first and third differences give the *range* of differences observed in this experiment). The middle difference is the largest of the remaining differences (based on the absolute value).

3 Diagnostic Experiments

Tables 3 lists the mean scores of the following 4 diagnostic runs for each language:

"words": The word-based index was used.

"words+qw": A second word-based index was used for which we stopped the indexing of common question words that we found in the training topics (note that the test topics were not inspected). The identified common words in questions were as follows:

- Simplified Chinese: 是, 什么, 谁, 事件, 关系, 列举, 请, 何谓, 有关, 案.
- Traditional Chinese: 什麼, 誰, 請, 請問, 何, 何謂.
- Japanese: です, 何, どの, よう, て, 関係, どんな, くださる, もの, 誰, つて, 教える, つく, 列挙, こと, 事, 出来, 事例, 人, 人物, よる, ん, 間, どう, どういう, 事件, 起きる, 関連.

"words+qw+bf": Blind feedback run (50% based on the "words+qw" run and 50% based on the first 3 rows retrieved by the "words+qw" run).

"ngram": The n-gram index was used.

3.1 Removing Question Words

Tables 4 and 5 isolate the impact of removing common question words (the "words+qw" score minus the "words" score, full relevance).

Table 4 shows that removing question words led to a statistically significant increase in mean GenS10 for 2 of the 3 languages and in mean average precision for all 3 languages.

Table 5 shows that removing question words had large impacts on some individual questions, including positive GenS10 differences at and exceeding 0.50 which indicate that answer documents moved into the first 10 documents retrieved for some topics.

3.2 Blind Feedback

Tables 6 and 7 isolate the impact of the blind feedback technique (the "words+qw+bf" score minus the "words+qw" score, full relevance).

Table 6 shows that blind feedback led to a statistically significant increase in mean average precision for 2 of the 3 languages. Normally we would expect a decline in mean GenS10 from blind feedback (like we found last time in NTCIR-6 [10] and in 7 other groups'

Table 4. Mean Impact of Removing Question Words

Expt	ΔGS10 (95% Conf)	vs.
CS-qw	0.051 (0.017, 0.084)	26-8-63
CT-qw	0.025 (−0.003, 0.053)	17-8-69
JA-qw	0.055 (0.016, 0.094)	26-17-54
	ΔMAP	
CS-qw	0.066 (0.042, 0.091)	75-14-8
CT-qw	0.019 (0.008, 0.030)	45-12-37
JA-qw	0.082 (0.056, 0.109)	76-17-4

Table 5. Per-Topic Impact of Removing Question Words

Expt	3 Extreme GS10 Diffs (Topic)
CS-qw	0.82 (376), 0.77 (379), −0.29 (366)
CT-qw	0.83 (397), 0.50 (404), −0.48 (396)
JA-qw	0.99 (19), 0.71 (291), −0.40 (137)
	3 Extreme AP Diffs (Topic)
CS-qw	0.62 (381), 0.46 (379), −0.28 (93)
CT-qw	0.26 (197), 0.17 (402), −0.11 (387)
JA-qw	0.49 (249), 0.47 (37), −0.37 (110)

Table 6. Mean Impact of Blind Feedback

Expt	ΔGS10 (95% Conf)	vs.
CS-bf	−0.002 (−0.022, 0.017)	16-14-67
CT-bf	0.003 (−0.021, 0.027)	20-19-55
JA-bf	0.011 (0.000, 0.023)	14-6-77
	ΔMAP	
CS-bf	0.014 (−0.010, 0.038)	45-50-2
CT-bf	0.025 (0.004, 0.045)	55-38-1
JA-bf	0.033 (0.013, 0.054)	61-32-4

Table 7. Per-Topic Impact of Blind Feedback

Expt	3 Extreme GS10 Diffs (Topic)
CS-bf	−0.48 (55), −0.31 (83), 0.27 (74)
CT-bf	0.63 (444), 0.32 (396), −0.37 (442)
JA-bf	0.27 (137), 0.21 (297), −0.15 (158)
	3 Extreme AP Diffs (Topic)
CS-bf	0.67 (376), 0.50 (379), −0.34 (93)
CT-bf	0.37 (448), 0.29 (179), −0.36 (395)
JA-bf	0.42 (110), 0.41 (154), −0.17 (249)

Table 8. Mean Impact of N-gram Parsing

Expt	ΔGS10 (95% Conf)	vs.
CS-ng	0.026 (−0.008, 0.060)	26-19-52
CT-ng	0.035 (−0.014, 0.083)	27-24-43
JA-ng	−0.097 (−0.159,−0.036)	15-35-47
	ΔMAP	
CS-ng	−0.000 (−0.028, 0.027)	52-45-0
CT-ng	−0.006 (−0.035, 0.024)	37-57-0
JA-ng	−0.100 (−0.150,−0.051)	27-70-0

Table 9. Per-Topic Impact of N-gram Parsing

Expt	3 Extreme GS10 Diffs (Topic)
CS-ng	0.68 (78), 0.63 (56), −0.53 (73)
CT-ng	0.97 (397), 0.93 (196), −0.57 (415)
JA-ng	−1.00 (109), −0.99 (105), 1.00 (240)
	3 Extreme AP Diffs (Topic)
CS-ng	−0.56 (93), −0.43 (317), 0.36 (56)
CT-ng	−0.42 (422), 0.33 (392), 0.40 (397)
JA-ng	−0.98 (105), −0.73 (109), 0.78 (240)

blind feedback systems (of the 2003 RIA workshop) in [9]), but it seems that the early success rate was high enough for this year's topics that relatively few detrimental documents were fed in, and the impact on mean GenS10 was fairly neutral.

[1] gives a theoretical explanation for why different retrieval approaches are superior when seeking just one relevant item instead of several. In particular, it finds that when seeking just one relevant item, it can theoretically be advantageous to use *negative* pseudo-relevance feedback to encourage more diversity in the results (i.e. after retrieving the first item, assume it is *not* relevant when deciding what to retrieve next; duplicate filtering is a special case of negative feedback).

3.3 Words vs. N-grams

Tables 8 and 9 isolate the impact of using overlapping n-grams instead of words (the "ngram" score minus the "words" score, full relevance).

Table 8 shows that using n-grams instead of words led to a statistically significant decline in both mean GenS10 and mean average precision for Japanese.

Table 9 shows that word and n-gram techniques strongly favor different topics for each language, including large impacts on the rank of the first answer document retrieved as per the large swings in the GenS10 score for some topics. For example, n-grams scored much lower than words in topic ACLIA1-JA-T109 (ラスカー賞とノーベル賞とはどういう

ものでゝどういう関係があるのか知りたいです。(What are the Lasker Awards and Nobel Prize, and what relationship do they have to each other?)). We haven't had time to fully investigate, but the n-gram approach generates a lot more terms and we suspect it may be harder for the key terms (Lasker Awards, Nobel Prize) to stand out in n-gram mode when there are a lot of other words in the question. (Common question words were not removed for either the "ngram" or "words" runs because the "ngram" mode cannot identify words.)

4 Precision to Depth 3000

One of our submitted runs for each language (hereinafter called the '01 run') was actually a depth probe run from sampling the plain "words" run for the language.

The base "words" run was retrieved to depth 10000 for each topic. The first 100 rows of the submitted 01 run contained the following rows of the base run in the following order:

```
1, 2, ..., 10,
20, 30, ..., 100,
200, 300, ..., 1000,
2000, 3000, ..., 10000,
15, 25, ..., 95,
150, 250, ..., 950,
1500, 2500, ..., 9500,
125, 175, ..., 975,
1250, 1750, ..., 9750.
```

The remainder of the 01 run was the leftover rows from the base run until 1000 had been retrieved (rows 11, 12, 13, 14, 16, ..., 962).

This ordering (e.g. depth 10000 before depth 15) was chosen because of uncertainty of how deep the judging would be. As long as the top-37 were judged, we would have sampling to depth 10000; the extra sample points would just improve the accuracy. The 01 run was given highest precedence for judging. It turned out that only at least the top-30 were judged for all topics, but this was still enough to give us sampling to depth 3000.

Tables 10, 11 and 12 show the results of the sampling for each language. The columns are as follows:

- "Samples": The depths of the sample points for this depth range. The 7 depth ranges are 1-5, 6-10, 11-50, 51-100, 101-500, 501-1000 and 1001-3000. The samples for each depth range are always uniformly spaced, and they always end at the last point of the depth range being sampled. The total number of sample points (over the 7 rows of the table) adds to 30 for all 3 languages.

Table 10. Marginal Precision of Simplified Chinese Base Run at Various Depths

Samples	Precision	EstRel/Topic
1, 2, ..., 5	0.542	2.7
6, 7, ..., 10	0.462	2.3
20, 30, ..., 50	0.294	11.8
60, 70, ..., 100	0.181	9.1
200, 300, ..., 500	0.080	32.0
600, 700, ..., 1000	0.033	16.5
2000, 3000	0.005	10.3
		(w/ partial)
1, 2, ..., 5	0.784	3.9
6, 7, ..., 10	0.726	3.6
20, 30, ..., 50	0.541	21.6
60, 70, ..., 100	0.410	20.5
200, 300, ..., 500	0.175	70.1
600, 700, ..., 1000	0.072	36.1
2000, 3000	0.026	51.5

- "Precision": Estimated precision over the depth range (based on dividing the number of relevant documents among the samples (over all topics) by the total number of samples).

- "EstRel/Topic": Estimated number of relevant items retrieved per topic for the sampled depth range. This is the Precision multiplied by the size of the depth range.

Because each sample point is at the deep end of the range of rows it represents, the sampling should tend to underestimate precision for each depth range (assuming that precision tends to fall with depth, which appears to be the case for all 3 languages).

Table 13 shows the sums of the estimated number of relevant items per topic over all depth ranges in its first row. The official number of relevant items per topic for each language is listed in the second row. The final row of the table just divides the official number of relevant items by the estimated number in the first 3000 retrieved (e.g. for Simplified Chinese (CS), 55.2/84.6=65%). This number should tend to be an overestimate of the percentage of all relevant items that are judged (on average per topic) because there may be relevant items that were not matched by the query in the first 3000 rows.

However, the sampling was very coarse at the deeper ranks, e.g. for Simplified Chinese, 1 relevant item out of 194 samples in the 1001-3000 range led to an estimate of 10.3 relevant items per topic in this range. If the sampling had turned up 0 or 2 relevant items, a minor difference, the estimate would have been 0 or 20.6 relevant items per topic in this range, leading to a substantially different sum (74.3 or 94.9

Table 11. Marginal Precision of Traditional Chinese Base Run at Various Depths

Samples	Precision	EstRel/Topic
1, 2, ..., 5	0.440	2.2
6, 7, ..., 10	0.343	1.7
20, 30, ..., 50	0.221	8.8
60, 70, ..., 100	0.123	6.2
200, 300, ..., 500	0.043	17.0
600, 700, ..., 1000	0.023	11.7
2000, 3000	0.016	31.9
		(w/ partial)
1, 2, ..., 5	0.674	3.4
6, 7, ..., 10	0.594	3.0
20, 30, ..., 50	0.421	16.8
60, 70, ..., 100	0.312	15.6
200, 300, ..., 500	0.108	43.2
600, 700, ..., 1000	0.051	25.3
2000, 3000	0.021	42.1

Table 12. Marginal Precision of Japanese Base Run at Various Depths

Samples	Precision	EstRel/Topic
1, 2, ..., 5	0.520	2.6
6, 7, ..., 10	0.416	2.1
20, 30, ..., 50	0.235	9.4
60, 70, ..., 100	0.146	7.3
200, 300, ..., 500	0.067	26.8
600, 700, ..., 1000	0.029	14.4
2000, 3000	0.021	41.2
		(w/ partial)
1, 2, ..., 5	0.755	3.8
6, 7, ..., 10	0.690	3.4
20, 30, ..., 50	0.469	18.8
60, 70, ..., 100	0.357	17.9
200, 300, ..., 500	0.148	59.2
600, 700, ..., 1000	0.076	37.8
2000, 3000	0.041	81.6

instead of 84.6). We should compute confidence intervals for these estimates, but have not yet done so. Also, there is a lot of variance across topics, which we have not had time to analyze for this paper (however, anyone with access to the run files could in principle carry out such an analysis themselves given the above information on how the sample points were chosen).

These preliminary estimates of judging coverage for the NTCIR-7 ACLIA IR4QA collections (65% for Simplified Chinese, 32% for Traditional Chinese, 41% for Japanese) tend to be lower than the estimates we produced for the NTCIR-6 CLIR collections last year (58% for Chinese (Traditional), 78% for Japanese, 100% for Korean) [10]. For comparision, we have also produced estimates (using similar approaches in other work) for some CLEF 2008 collections (55% for German, 52% for French, 53% for English, 25% for Persian) [11], some CLEF 2007 collections (55% for Czech, 69% for Bulgarian, 83% for Hungarian) [12] and some TREC 2006 collections (18% for TREC Legal and 36% for TREC Terabyte) [13].

Test collections can still be useful despite incomplete coverage. e.g. [14] found for depth-100 pooling on the old TREC collections of approximately 500,000 documents that "it is likely that at best 50%-70% of the relevant documents have been found; most of these unjudged relevant documents are for the 10 or so queries that already have the most known answers." Fortunately, [14] also found for such test collections that "overall they do indeed lead to reliable results."

Table 13. Estimated Percentage of Relevant Items that are Judged, Per Topic

	CS	CT	JA
Estimated Rel@3000	84.6	79.6	103.9
Official Rel	55.2	25.4	42.2
Percentage Judged	65%	32%	41%
Est. Rel+Par@3000	207.4	149.3	222.4
Official Rel+Par	97.8	55.4	86.8
Percentage Judged	47%	37%	39%

5 Comparing Full and Partial Relevance

The 5 submitted runs for each language actually were the 5 diagnostic runs described in the previous sections. The precedence codes were as follows:

```
01 depth probe
02 words+qw
03 ngram
04 words+qw+bf
05 words
```

Table 14. Mean Scores of Submitted Runs (Full Relevance and Partial Relevance)

Run	GenS10	S1	MAP
OT-CS-CS-01-T	0.870	60/97	0.291
OT-CS-CS-02-T	0.922	66/97	0.488
OT-CS-CS-03-T	0.897	59/97	0.421
OT-CS-CS-04-T	0.919	70/97	0.502
OT-CS-CS-05-T	0.871	60/97	0.422
OT-CT-CT-01-T	0.825	49/94	0.260
OT-CT-CT-02-T	0.850	50/94	0.385
OT-CT-CT-03-T	0.859	53/94	0.361
OT-CT-CT-04-T	0.853	52/94	0.410
OT-CT-CT-05-T	0.825	49/94	0.366
OT-JA-JA-01-T	0.867	57/97	0.307
OT-JA-JA-02-T	0.921	62/97	0.506
OT-JA-JA-03-T	0.769	48/97	0.323
OT-JA-JA-04-T	0.932	64/97	0.539
OT-JA-JA-05-T	0.866	57/97	0.423
	p-GenS10	p-S1	p-MAP
OT-CS-CS-01-T	0.950	76/97	0.370
OT-CS-CS-02-T	0.975	86/97	0.629
OT-CS-CS-03-T	0.976	80/97	0.566
OT-CS-CS-04-T	0.970	88/97	0.634
OT-CS-CS-05-T	0.950	76/97	0.564
OT-CT-CT-01-T	0.925	74/95	0.323
OT-CT-CT-02-T	0.940	74/95	0.511
OT-CT-CT-03-T	0.935	73/95	0.501
OT-CT-CT-04-T	0.936	75/95	0.552
OT-CT-CT-05-T	0.926	74/95	0.491
OT-JA-JA-01-T	0.962	80/98	0.389
OT-JA-JA-02-T	0.979	85/98	0.670
OT-JA-JA-03-T	0.869	68/98	0.425
OT-JA-JA-04-T	0.979	86/98	0.698
OT-JA-JA-05-T	0.959	80/98	0.566

Table 14 shows the mean scores for the submitted runs. The top-half of the table just counts the fully relevant documents as relevant, and the bottom-half counts both full and partially relevant documents. While the scores are higher when partially relevant documents are included, we don't see any substantial differences in the system rankings from whether full or partial relevance is used.

6 Conclusions

We conducted several experiments in finding documents containing the answers to natural language questions posed in Chinese and Japanese. For this task, a "relevant" document is one that answers the question.

In our sampling experiment, we found that, on average per topic, the percentage of relevant documents assessed was probably less than 65% for Simplified Chinese, 32% for Traditional Chinese and 41% for Japanese. However, test collections can still be useful despite incomplete coverage. Furthermore, our preferred measure for this task, Generalized Success@10 (GenS10), only looks at the rank of the first relevant document retrieved for each topic. One good document answering the question is all that a user needs for this task. Also, most topics were fully judged to at least depth-30 for all of our submitted runs. For GenS10, it's a good use of assessor resources to have relatively shallow judging (e.g. depth-30) to gain more test questions than usual (e.g. 94-97).

The step of removing common question words was found to substantially boost effectiveness. In particular, it produced statistically significant increases in mean GenS10 for some languages.

Blind feedback was not found to produce substantial declines in GenS10 in this year's topics, contrary to our findings on several other test collections. We suspect that the early success rate was high enough for these topics that bad documents were seldom fed in.

Indexing using overlapping n-grams instead of by segmenting into words was found to produce substantial differences in GenS10 on individual questions, some postive and some negative. Overall, the only statistically significant impact found for mean GenS10 was a decline from using n-grams for Japanese.

Broadening the definition of relevance to include partially relevant documents was not found to lead to any substantial differences in the rankings of our submitted runs.

The NTCIR-7 ACLIA IR4QA Task supports evaluation of retrieval techniques for an important retrieval scenario, seeking just one document to answer a question. Our experiments show that the choice of technique can have a substantial impact on the rank of the first answer document.

References

[1] H. Chen and D. R. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. *Proceedings of SIGIR 2006*, pp. 429-436.

[2] Cross-Language Evaluation Forum (CLEF) web site. http://www.clef-campaign.org/.

[3] A. Hodgson. Converting the Fulcrum Search Engine to Unicode. *Sixteenth International Unicode Conference*, 2000.

[4] NTCIR (NII Test Collection for IR Systems) Project Home Page. http://research.nii.ac.jp/ntcir/.

[5] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu and M. Gatford. Okapi at TREC-3. *Proceedings of TREC-3*, 1995.

[6] T. Sakai, N. Kando, Chuan-Jie Lin, T. Mitamura, D. Ji, Kuang-hua Chen and E. Nyberg. Overview of the NTCIR-7 ACLIA IR4QA Task. To appear in *Proceedings of NTCIR-7*, 2008.

[7] Text REtrieval Conference (TREC) Home Page. http://trec.nist.gov/.

[8] S. Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer^{TM} at CLEF 2005. *Working Notes for the CLEF 2005 Workshop*.

[9] S. Tomlinson. Early Precision Measures: Implications from the Downside of Blind Feedback. *Proceedings of SIGIR 2006*, pp. 705-706.

[10] S. Tomlinson. Sampling Precision to Depth 9000: Evaluation Experiments at NTCIR-6. *Proceedings of NTCIR-6*, 2007.

[11] S. Tomlinson. German, French, English and Persian Retrieval Experiments at CLEF 2008. *Working Notes for the CLEF 2008 Workshop*.

[12] S. Tomlinson. Sampling Precision to Depth 10000: Evaluation Experiments at CLEF 2007. *Working Notes for the CLEF 2007 Workshop*.

[13] S. Tomlinson. Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track. *Proceedings of TREC 2006*.

[14] J. Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments? *Proceedings of SIGIR'98*, pp. 307-314.