OASIS at NTCIR-3: Monolingual IR Task

Vitaliy KLUEV
The Core and Information Technology Center
The University of Aizu
Tsuruga, Ikki-machi, Aizu-Wakamatsu city, Fukushima, 965-8580, Japan
vkluev@u-aizu.ac.jp

Abstract

We participated in the monolingual Japanese and English information retrieval tasks. Results of our experiments using the OASIS system have been presented in this article. Our aim was to check the reliability of pseudo-relevance feedback. Our strategy attempted to select the best terms from the top ranked documents and to expand the initial user query using them. Four automatic runs were submitted for each task. Two of them were done using relevance feedback; and other two were carried out without any feedback. We gained nothing regarding retrieval improvement in our test with query expansion.
Keywords: *OASIS, search engine, phrasal indexing, vector space model, full text searching.*

1 Introduction

The OASIS system participated in the monolingual Japanese and English information retrieval tasks. The system was designed by the international team [4]. It is dedicated to search for text information in the Internet. The aim of our participatio was to check the quality of the search carried out using pseudo-relevance feedback: the number of actually relevant documents on the top of the list presented to the user. A number of research reported the promising results of information retrieval using pseudo-relevance feedback [2, 3]. There are several approaches to expand initial user query. Their descriptions can be found in the following citations [5, 6].

The paper is organized as follows. Methods used in our tests are described in section 2. Explanations related the Japanese Retrieval Task are presented in section 2.1. Discussions on the English Retrieval Task are put into section 2.2. Final remarks can be found in section 3.

Table 1. Official Runs

Run	Collection	Query expansion
OASIS-J-J-D-01	ntc-j-mai99.txt	Yes
OASIS-J-J-D-02	ntc-j-mai98.txt	Yes
OASIS-J-J-D-03	ntc-j-mai98.txt	No
OASIS-J-J-D-04	ntc-j-mai99.txt	No

2 Technique Description

2.1 Japanese Retrieval Task

The key parameters of our system have been presented below. They are: Index Unit, Index Technique, Index Structure, Query Unit, IR Model, Ranking Technique and Query Expansion Strategies. Table 1 and Table 2 describe the official runs and the technique used.

To define the aforementioned parameters (see the item Query Expansion) we made several sets of tests using DryRun topics. Relevant documents corresponded these topics were given by the Executive Committee. Table 3 describes key parameters used in query expansion and their values utilized in the formal runs. The aim of using virtual word collocations is to catch context more accurately.

Participants were requested to present up to 1000 documents in response to each query. Because a number of relevant documents for each query does not exceed 1000 it was interesting to know how often the system retrieves all relevant items. Table 4 answers this question. It presents the effectiveness of retrieval. This table shows the best and the worst results of each run. Runs with query expansion produced the relatively poorer outcome. The string entitled ”all” presents the query numbers for which all relevant documents were retrieved. On the other hand, the ”zero” string indicates queries with nothing-relevant retrieval.

Figure 1 presents average results of the search. Tests using relevance feedback and query expansion produced worse results. We put data in our table ac-

Table 2. System description

Parameter	Description
Index Unit and Index Technique	Combination of bi-words and phrases. Overlapped bi-gram words were selected from every indexing document. Phrases (virtual word collocations) consisted up to 4 Japanese characters were automatically determined. Hiragana characters were used as word boundaries. Katakana sequences were considered as words. Hiragana characters were discarded.
Index Structure	Inverted index
Query Unit	The same as the Index Unit: combination of bi-words and phrases.
Query Method	Automatic
IR Model	Vector Space Model
Ranking	TF*IDF
Query Expansion	Query expansion was used only in runs OASIS-J-J-D-01 OASIS-J-J-D-02. Every query was processed twice. The first search generated 2 documents. Terms consisted of 4 characters were considered as candidates for expansion. These terms can be considered as a pseudo two-word collocation if we try to compare with English. Words that occurred four times were selected from this set. Their number did not exceed a half of the words in the original query. In the case of necessary, the random selection was utilized. Weights of all terms were corrected with a damping factor. It is set to 1/8.

Table 3. Parameters and their range used in Query Expansion

Parameter	Range tested	Values used in official runs
Length of virtual word collocation	a) 1 - 4 characters; b) more than three characters (only word collocations have been taken)	4 characters
Number of documents retrieved to expand queries	1 - 5	2
Number of terms to add to queries	a) 10 most heaviest; b) all terms the number of terms is equal to the number of terms in the original query	A half of the query length
Damping factor	1/16, 1/8, 1/4, 1/2 and 1	1/8
Threshold to select terms to expand queries (occurrence number)	1, 2, 3, 4 and any	3 - 4

Table 4. The best and worst retrieval results

	OASIS-J-J-D-04		OASIS-J-J-D-01		OASIS-J-J-D-03		OASIS-J-J-D-02	
	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax
all	5, 8, 10, 15, 16, 17, 22, 24, 25, 27, 28, 29, 32, 33 37, 40 43, 50	5, 8, 10, 15, 16, 24, 25, 29, 30, 32, 33, 37, 43, 50	8, 29, 32, 33, 40, 43	8, 32, 33,	2, 5, 7, 8, 10 , 15, 17, 20, 22, 24, 25, 26, 27, 29, 31, 38, 39, 40, 43, 44, 45, 47, 50	2,5, 7, 10, 15, 22, 24, 25, 26, 27, 28, 31, 38, 39, 40, 44, 47, 50	8, 17, 31, 40, 43, 44, 50	31, 40 44, 50
Number of queries	18	14	6	3	23	18	7	4
zero	47 (from 2 docs)	none	15 (from 3), 22 (from 1) 37 (from 8) 42 (from 4) 47 (from 2)	30 (from 2)	4 (from 2)	none	2 (from 1) 4 (from 2) 7 (from 4) 20 (from 1) 24 (from 5) 26 (from 3) 47 (from 1)	28 (1)
Number of queries	1	none	5	1	1	none	7	1

cording ”Rigid” results. This result surprised us. Outcome from this: indexing strategy using variable gram (up to four in our case) technique is good enough. But query expansion on the basis a long gram approach (four gram) produced the significantly worse results.

Figure 2 shows how many relevant hits were among the top ten retrieved documents presented to the user. This parameter is important for information retrieval systems because users usually want to see only the first page with retrieval results. As it was shown, outcome depends on queries. On overage, precision at 10 documents is equal to 0.29021. The method using pseudo-relevance feedback produced the better results only for questions 10, 41, 42 and 46.

Our system produced very poor results in response to query 47. The reason for this is as follows: we did not use any natural language processing techniques in our research to expand initial user queries. This is a common weakness of vector space model: if a query has been expressed in a special way using words or terms which are not common for a topic of interest then the search results are very far from expected ones. To improve the accuracy of the search any system has to define the topic or context of the query. After that it will be easier to expand the query using synonym dictionaries or thesauruses. An overage query length in these tests was about 10 terms (a virtual word or word collocation consisting of 1, 2, 3 or 4 characters). How can the search system predict (guess) the topic without human intervention? This question is still unanswered. Another unanswered question is about the number of new terms to expand the query. It is not clear how to select them, and how to add to the query. Is there a simple and at the same time effective method of indexing and searching for Japanese? We vote for simplicity. Variable (one- four) gram indexing method for Japanese is rough-and-ready compared to a complicated natural language processing approach. The way of the topic detection has to be connected with a thesaurus structure.

The aforementioned ideas have been connected very well with the OASIS approach. It can be characterized as follows: any OASIS server can create and support one or more topic specific indices. A special component (one in the whole system) keeps description of each topic. When one server receives a query from the end user, the aforementioned component assists to determine the query context. After that user query propagates to the small set of indices which could contain requested information. This style of searching is promising because the system can avoid retrieving garbage.

Table 5. Official Runs

Run	Collection	Query expansion
OASIS-E-E-D-01	ntc-e02-mai98.txt	No
OASIS-E-E-D-02	ntc-e02-mai98.txt	Yes
OASIS-E-E-D-03	ntc-e02-mai99.txt	No
OASIS-E-E-D-04	ntc-e02-mai99.txt	Yes

2.2 English Retrieval Task

The key parameters of our system and runs have been presented in Table 5 and Table 6.

Results of the search using pseudo relevance feedback are worse compared to the results obtained without query expansion. Figure 3 illustrates this outcome. There is a note: We could not test this approach because the English version of training topics were not available.

Table 7 shows the best and worst retrieval results. Precision at 10 documents has been presented in Figure 3. Average results of the search can be seen in Figure 4. The comparison in accuracy between Japanese and English retrieval can be done using Figure 5: Our system produced the better retrieval with the English test collections. In the case of English retrieval, runs without query expansion generated better results. Only queries 33 and 43 belong to the exceptions from this outcome.

3 Conclusions

Our tests showed that the search using pseudo-relevance feedback and query expansion produce relatively worse results compared to the search without query expansion. We gained nothing regarding retrieval improvement. Our results (parameters for query expansion) are strictly empirical. The approach used here is faulty. Its main idea was to select long-gram (four gram) terms to extend the initial query for the Japanese task and words, which occur fixed time in relevant documents for the English task. This feedback added only a noise. From this we need to study more how to select terms and parameters for query expansion. As we can see from Table 7, retrieval results for the English task are much better compared to the Japanese task. One of the reasons for this is to discard stop words from English texts.

How to decrease retrieval non-relevant documents using statistical methods? We believe that results of the search can be improved if test collections divide into several narrow topics related sets. Clustering methods will be utilized to divide test collections. Some approaches to select appropriate sets should be tested. We are planning to conduct corresponding experiments using Japanese and English test collections.

References

[1] Smart (a list of stopwords. ftp://ftp.cs.cornell.edu/pub/smart.

[2] *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*. National Center for Science Information Systems, Tokyo, Japan, 1999. (ISBN: 4-92-4600-77-6).

[3] *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*. National Institute of Informatics, Tokyo, Japan, 2001. (ISBN: 4-92-4600-96-2).

[4] A. Patel, L. Petrosjan and W. Rosenstiel, editor. *OASIS: Distributed Search System in the Internet*. St. Petersburg State University Published Press, St. Petersburg, Russia, 1999. (ISBN: 5-7997-0138-0).

[5] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. the ACM Press, 1999. ISBN: 0-201-39829-X.

[6] Tetsuya Sakai, Stephen E. Robertson and Stephen Walker. Flexible pseudo - relevance feedback for ntcir-2. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, 2001. (ISBN: 4-92-4600-96-2).

Table 6. System description

Parameter	Description
Index Unit and Index Technique	The index unit is a word. Stop words according to the standard de facto set [1] were eliminated.
Index Structure	Inverted index
Query Unit	The same as the Index Unit.
Query Method	Automatic
IR Model	Vector Space Model
Ranking	TF*IDF
Query Expansion	Query expansion was used only in runs OASIS-E-E-D-02 OASIS-E-E-D-04. Every query was processed twice. The first search generated 2 documents. Words that occurred four times were considered as candidates for expansion. Their number did not exceed a half of the words in the original query. In the case of necessary, the random selection was utilized. Weights of all terms were corrected with a damping factor. It is set to 1/8. We used the same parameters as for the Japanese retrieval task.

Table 7. The best and worst retrieval results

	OASIS-E-E-D-01		OASIS-E-E-D-02		OASIS-E-E-D-03		OASIS-E-E-D-04	
	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax
all	5, 12, 19, 23 24, 29, 32, 33, 34, 35, 36, 37 38, 42, 50	5, 12, 19, 23, 24, 29, 32, 33, 34, 35, 36, 37, 38, 39, 42, 43, 45, 50	29, 33, 37, 43,	29, 37, 43	5, 14, 19, 20 , 21, 23 24, 26, 28, 29, 31, 33, 34, 35, 37, 39, 45, 46, 50	2, 4, 5, 14, 19, 20, 21, 23, 24, 26, 27, 28, 29, 31, 33, 34, 35, 39, 42, 43, 45, 46, 50	24, 29, 45, 46, 50	2, 27, 29, 43, 45, 46
Number of queries	17	18	4	3	19	23	5	6
zero	none	none	5 (from 2), 24 (from 1) 34 (from 2) 38 (from 3) 50 (from 1)	3 (from 3) 5 (from 5) 24 (from 1) 34 (from 2) 39 (from 1) 50 (from 1)	none	none	14 (from 2) 31 (from 1) 33 (from 1) 34 (from 1)	14 (from 3) 31 (from 1) 33 (from 1) 42 (from 1)
Number of queries	none	none	5	6	none	none	4	4

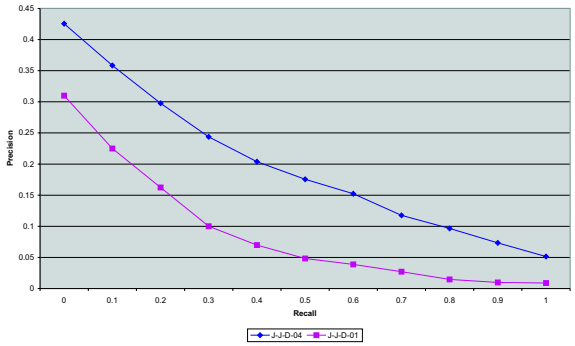

Figure 1. Japanese monolingual task

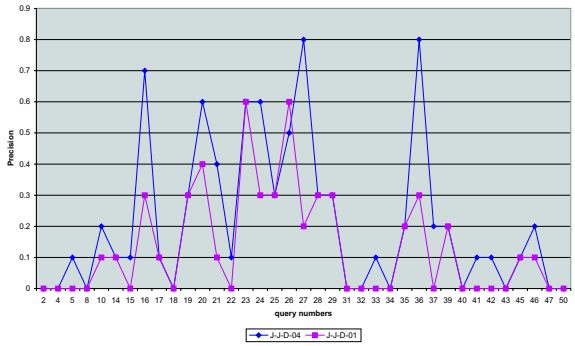

Figure 2. Japanese monolingual task: Precision at 10 docs

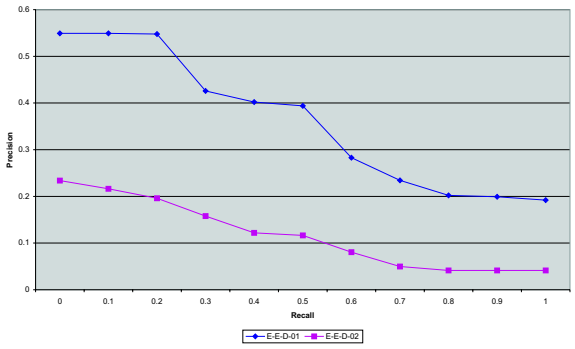

Figure 3. English monolingual task

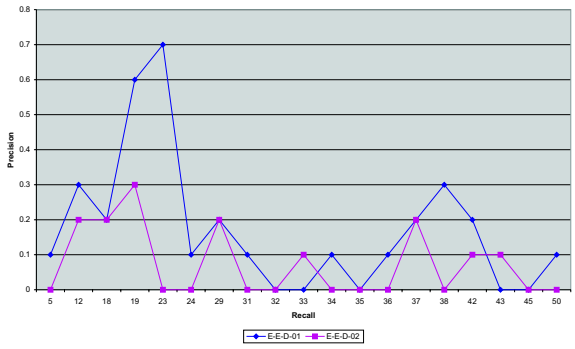

Figure 4. English monolingual task: Precision at 10 docs

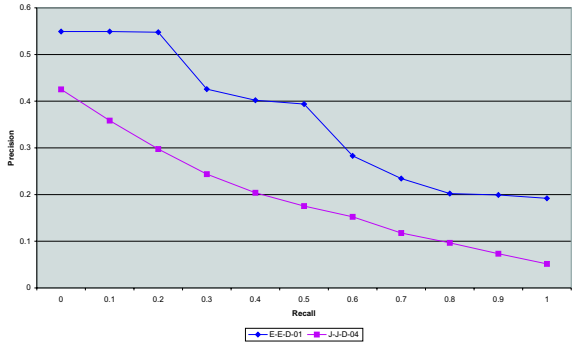

Figure 5. English and Japanese monolingual tasks