Notes on Phrasal Indexing
JSCB Evaluation Experiments at NTCIR AD HOC

Sumio FUJITA
JUSTSYSTEM Corporation
Brains park, Tokushima, 771-0189 JAPAN
+81-88-666-1000
Sumio_Fujita@justsystem.co.jp

ABSTRACT

The evaluation experiments of the JSCB team are described with a focus on noun phrase indexing and its weighting issues in ad hoc text retrieval.

Experiments on the effects of supplemental noun phrase indexing in view of the effect of various length of queries are reported.

The results show that the noun phrase indexing outperforms single word only indexing with long queries while single word only indexing performs slightly better with short queries.

A new weighting method for phrasal terms is also evaluated and improvement is observed.

Keywords

Phrasal indexing, noun phrase indexing, phrasal terms, weighting, vector space model.

1. INTRODUCTION

Automatic indexing of modern information retrieval systems typically adopts bag-of-word representation, in which each word is considered as a dimension of the vector representing an information item, as internal representation of "aboutness". It is well known that such simple representation usually performs, as well as, if not better than, some more sophisticated ones according to empirical evaluations.

Grammatical relations or functional words are normally considered as neutral in view of thematic discrimination of text documents. On the other hand, content words(or lexemes, if we need to be more attentive for linguistic terminology) are semiologically meaningful units in language systems which refer to conceptual/substantial entities or relations in the subject domain described by the documents. It is plausible that the author of documents and the user submitting search requests share the same terminology when describing the subject concept in question either in their documents or in queries.

As indicated by one of the earliest pioneers in modern information retrieval research [1], a word is sometimes ambiguous and too vague to be used as indexing unit in isolation, in other words, single words are not very meaningful unit in specific domain terminology, but no more than a part of larger units i.e. compound words, which are more meaningful information carriers in the domain terminology. Some adjectives in technical English are too vague and useless for indexing terms by themselves, but they are able to compose good discriminators like *high* in *high subsonic speeds*, *high aspect ratio*, *higher order logic*, *higher dimension* …

Another problem with single word indexing is the possible loss of information by ignoring local syntactic relations in noun phrases.

Using only single word terms "情報"(information) and "管理"(management), the system could not distinguish "情報管理"(information management) from "管理情報"(information for management).

Especially in Japanese language, such noun compounds sometimes make domain specific terminology that is usually useful as a good discriminator of subject concept description.

Given such terminological characteristics, indexers introduced precoordination of indexing terms mainly adopting phrasal terms in order to preserve syntactic relations.

In automatic indexing, such phrasal terms are extracted either by statistic or linguistic procedures [7].

Although the idea behind it is reasonable, no reliable performance improvement by using such phrasal terms was reported [4][3] until the TREC conference, that enabled researchers to carry out experiments with large scale English test collections, started in 1992.

Some positive results for phrasal indexing were reported in the past TREC experiments by a few sites [8][11][12], especially in TREC5, where NLP track concentrated on evaluating effects of various NLP techniques for retrieval [9].

Among them, Strzalkowski et al. suggested that the use of syntactic phrases was more effective with longer queries [10].

On the other hand, Lewis reported that simple noun phrase based classifier performed better than single word based classifier, when much larger number of terms are used [5].

Given such observations found in the literature, we hypothesize that the reasons why the effects of phrasal terms in text retrieval are considered as inconsistent and uncertain, are not only because they are sensitive to characteristics and quality of phrases extracted, but also because of distribution

characteristics of phrasal terms that is sensitive to collection size and query length.

2. PHRASAL INDEXING

Our approach consists of utilizing noun phrases extracted by linguistic processing as supplementary indexing terms in addition to single word terms contained in phrases. Phrases and constituent single terms are treated in the same way, both as independent terms, where the frequency of each term is counted independently based on its occurrences .

It is not clear whether such a weighting scheme is adequate, since phrases tend to have a lower frequency, consequently a higher weight as indicated in [12].

Phrases may have such a low frequency that they rarely match, consequently they are not so useful. They are probably also over-weighted such that they would cause some bias in scoring when matched. The following example illustrates this point.

For the description field of the NTCIR topic 35,

「分散環境における電子図書館についての研究はないか」

(Is there any research regarding electronic libraries in distributed environments?)

The term extraction subsystem extracts the following single words (assume that the word 「研究」 (research) is a stopword.):

「分散」(distribution),「環境」(environment),「電子」(electronic),「図書館」(library)

And also, the system detects the following noun phrases by applying linguistic rules:

「分散-環境」,「電子-図書館」

All terms are weighted by term frequency 1 of query part, term frequency of document part and their own IDF value computed against the target database.

Retrieval status value(RSV) is computed by a linear combination of each term weight as follows:

w(分散)+w(環境)+w(分散-環境)+w(電子)+w(図書館)+w(電子-図書館)

Since Japanese written language provides no explicit word boundary marker, single words are defined solely on the basis of dictionary entry, and there are alternative possibilities of term extraction depending on the system.

The problem arises if the term extraction system outputs the following terms:

「分散」(distribution),「環境」(environment),「電子図書館」(electronic library)

RSV is as follows:

w(分散)+w(環境)+w(分散-環境)+w(電子図書館)

Intuitively, the weights are “doubly added” for the first noun phrase part which may cause a bias in retrieval results favoring the documents containing terms of the first noun phrase.

In fact, such a weighting scheme might heavily violate term independence assumption of vector space retrieval.

Another problem concerning phrasal indexing is related to processing noisy or corrupted data. Since in phrasal indexing miss-match in one word is propagated to phrase levels, performance is more sensitive to noises than in single word indexing. In this respect, supplemental phrasal indexing with single words is better than using only precoordinated longer phrases.

In order to address such a problem, we adopted down-weighting of term coefficient for phrasal terms which simply means a decreasing phrasal term weight against single word terms.

3. SYSTEM DESCRIPTION

For the JSCB NTCIR experiments, we used the engine of Justsystem ConceptBase Search™ version 1.2 as the base system.

Two Pentium II™ machines (450MHz) running Windows NT™ 4.0 with 384MB memory and 9 GB hard disk are used for experiments.

The document collections are indexed wholly automatically, and converted to inverted index files of terms.

3.1 Term Extraction

Queries and documents in target databases are analyzed by the same module that decomposes an input text stream into a word stream and parses it using simple linguistic rules , in order to compose possible noun phrases.

Extracted units are single word nouns and noun equivalent words as well as simple linguistic noun phrases which consist of a sequence of nouns.

3.2 Vector Space Retrieval

Each document is represented as a vector of weighted terms by tf*idf in inverted index files and the query is converted in similar ways [7].

Similarity between vectors representing a query and documents are computed using the dot-product measure, and documents are ranked according to decreasing order of RSV.

3.3 Automatic Feedback

Automatic feedback strategy using pseudo-relevant documents is adopted for automatic query expansion.

The system submits the first query generated automatically from topic descriptions against the target document database, and considers the top n documents from relevant ranking list returned as relevant.

The term selection module extracts salient terms from these pseudo-relevant documents and adds them to the query vector.

Then the expanded query vector is submitted against the target database again and the final relevance ranking is obtained.

3.4 Relevance Feedback

Given users' relevance judgement against retrieved documents by the first search, the system can use these documents as positive examples. In the manual run JSCB3, we adopted traditional user relevance feedback to create final queries.

The procedure is exactly the same as automatic feedback, where top n documents are used, except example documents are viewed and judged by a user.

Since, in real situation of system usage, user relevance feedback is much more likely to be utilized frequently than automatic feedback, it is interesting to compare these two feedback procedures by simulating a real retrieval situation.

3.5 Term Selection

The following three term selection measures described in [6], are applied for Japanese text retrieval and evaluated in pre-test experiments.

Each term in example documents are scored by one of the following weighting functions.

1)CLARIT™ Thesaurus Discovery

This method involves some term frequency and document frequency based heuristics measures described in [2].

2)Rocchio

Standard Rocchio formula is also used in manual runs.

$$w(t) = IDF(t) * \frac{\sum_{d \in D} TF_d(t)}{|D|} \quad (1)$$

where D: example document set

|D|: number of documents in D

t: term

3)CLARIT Probabilistic term weighting

Although it is not used in the official submission runs, we evaluated a revised version of the standard Robertson-Sparck Jones formula described in [6].

The terms thus scored are sorted in decreasing order of each score and cut off at a threshold determined empirically.

In effect, the following parameters in feedback procedures should be decided:

1)How many documents to be used for feedback?

2)Which function to use to score terms?

3)Where to cut off ranked terms?

4)How to weight these additional terms?

These parameters are carefully adjusted using pre-test queries (topic 1-30) and their relevance judgement provided by NTCIR organizer.

4. EXPERIMENTS

We submitted three official results, corresponding to automatic short query (JSCB1), automatic long query (JSCB2) and manual (JSCB3).

The experiments are designed to measure effects of phrasal term indexing regarding different length of queries and different weighting.

For each run, our basic strategy is as follows:

1) To add terms automatically to the query from possible sources like each field in topic description("description" in JSCB1, all fields in JSCB2,JSCB3), automatic feedback procedure(JSCB1,JSCB2, indicated "AFB" in tables) and relevance feedback procedure(JSCB3, indicated "RFB" in tables);
2) To optimize the weighting coefficient for each term set from the different sources according to the reliability of the source.

Coefficient parameters are adjusted in pre-test experiments and fixed until the final submission runs. Thus, three parameter sets are decided:

JSCB1 set: optimized for short query and automatic feedback,

JSCB2 set: optimized for long query and automatic feedback,

JSCB3 set: optimized for long query and interactive relevance feedback.

In the following experiments, one of these three parameter sets is always applied where applicable.

As the official relevance assessment file, rl-je1_v001.txt (JE-1 judgement) is mainly used.

4.1 Automatic Short Query Run

This is a compulsory run for the site submitting automatic runs, using only "description" fields of topic description.

The following runs are examined:

1. Single word and phrasal terms with down-weighting for phrases and automatic feedback(JSCB1)

Run description	AFB	Avg. Prec	R-Prec
Phrasal terms with down-weighting (JSCB1)	Yes	3596	3505
Phrasal terms with down-weighting	No	3227	3341
Word terms(JSCB1-WD)	Yes	3621	3702
Word terms	No	3230	3465
Phrasal terms with normal tf*idf	Yes	3277	3204
Phrasal terms with normal tf*idf	No	2893	2914

Table 1: Performance in Automatic Short Experiments (JE-1)

2. Single word and phrasal terms with down-weighting for phrases and no automatic feedback
3. Single word terms with automatic feedback(JSCB1-WD)
4. Single word terms with no automatic feedback
5. Single word and phrasal terms with normal tf*idf weighting and automatic feedback
6. Single word and phrasal terms with normal tf*idf weighting and no automatic feedback

Since in the following experiments, phrasal terms are always used with their constituent single words, phrasal term run generally means single word and phrasal term run by default.

Run	JSCB1	JSCB1-WD	JSCB1-COMB
Retrieved:	53000	53000	53000
Relevant:	1910	1910	1910
Rel_ret:	1374	1363	1377
Interpolated Recall Precision Averages			
At 0.00	0.6834	0.7024	0.6872
At 0.10	0.6172	0.6346	0.6316
At 0.20	0.5209	0.5788	0.5698
At 0.30	0.4820	0.5026	0.5143
At 0.40	0.4131	0.4332	0.4355
At 0.50	0.3731	0.3594	0.3756
At 0.60	0.3148	0.2928	0.3067
At 0.70	0.2565	0.2501	0.2667
At 0.80	0.2203	0.1992	0.2190
At 0.90	0.1437	0.1229	0.1393
At 1.00	0.0923	0.0811	0.0875
Average precision	**0.3596**	**0.3621**	**0.3715**
At 5 docs:	0.4528	0.5019	0.4792
At 10 docs:	0.3962	0.4245	0.4283
At 15 docs:	0.3572	0.3572	0.3736
At 20 docs:	0.3245	0.3198	0.3340
At 30 docs:	0.2843	0.2811	0.2881
At 100 docs:	0.1455	0.1398	0.1428
At 200 docs:	0.0919	0.0887	0.0909
At 500 docs:	0.0462	0.0460	0.0459
At 1000 docs:	0.0259	0.0257	0.0260
R-Precision Exact:	**0.3505**	**0.3702**	**0.3780**

Table 2:Detailed Performance of JSCB1, JSCB1-WD and JSCB1-COMB(JE-1)

Since initial queries are short (in average, 6.2 single word terms and 1.9 phrasal terms) and they do not contain enough terms, the automatic feedback procedure contributes to 11% to 13 % of consistent improvements in average precision in all cases.

The final queries contain 33.7 single words and 18.1 phrases in average.

The phrasal term down-weighting run (JSCB1) outperformed normal tf*idf run with about 10% to 12 % improvement in avg. precision but the single word only run (JSCB1-WD) was still slightly better.

Although the phrasal term run (JSCB1) performed better for pre-test queries (1-30), it is not the case with the test queries (31-83) as shown in Table 1.

Firstly, we suspected that there might be a possible bias in the judgement pooling since many of the run submitted are word based, the pooling might be favored for word based retrieval but it was not the case.

For the JSCB1, 25421 retrieved documents were unjudged while 26645 for JSCB1-WD were unjudged. In fact, the pooling was favored for the JSCB1 run rather than JSCB1-WD.

Secondly, we thought that the down-weighting parameters were not enough and we tried various parameters. But even largely down-weighted, phrasal term run could not perform better than JSCB1-WD.

The effect of phrasal term indexing is not really clear in these experiments. The following precise results of JSCB1 and JSCB1-WD show that the word term run is rather precision (or initial precision) favored while the phrasal term run retrieved more relevant documents.

As a post-submission experiment, we planned a combination run using single word based initial query and phrasal term based final query, which we call JSCB1-COMB, hopefully taking advantage of both higher initial precision of JSCB1-WD and better recall of JSCB1. As Table 2 shows, this approach is promising, and we get a generally better result than the official submission run.

4.2 Automatic Long Query Run

The second official run that we named JSCB2 uses automatic query construction from all fields in topic description.

Since NTCIR topic descriptions are very rich in terms contained, it is important to adjust weighting for each term according to its importance in the description.

We adjusted term weights according to the fields in which the term appeared since shorter fields seem to describe more concentrated information than longer fields.

Topic fields used	AFB	Avg. Prec	R-Prec
<title>,<description>,<narrative>,<concepts> (=JSCB2)	Yes	4436	4301
<title>,<description>,<narrative>,<concepts>	No	4380	4309
<title>,<description>,<narrative>	Yes	4085	4113
<title>,<description>,<narrative>	No	3974	4063
<title>,<description>,<concepts>	Yes	4262	4201
<title>,<description>,<concepts>	No	4157	4176
<title>,<description>	Yes	3611	3761
<title>,<description>	No	3393	3585
<description>[1] (JSCB1)	Yes	3495 (3596)	3639 (3505)
<description>[1] (JSCB1-no AFB)	No	3305 (3227)	3442 (3341)
<title>	Yes	2779	2953
<title>	No	2475	2735

Table 4: Performance using different query fields with phrasal terms(JE-1)

Topic fields used	AFB	Avg. Prec	R-Prec
<title>,<description>,<narrative>,<concepts> (=JSCB2-WD)	Yes	4184	4235
<title>,<description>,<narrative>,<concepts>	No	4166	4227
<title>,<description>,<narrative>	Yes	3948	3986
<title>,<description>,<narrative>	No	3896	4006
<title>,<description>,<concepts>	Yes	4034	4100
<title>,<description>,<concepts>	No	4014	4049
<title>,<description>	Yes	3536	3650
<title>,<description>	No	3406	3613
<description>[1] (JSCB1-WD)	Yes	3381 (3621)	3650 (3702)
<description>	No	3230	3465
<title>	Yes	2797	2983
<title>	No	2573	2774

Table 5: Performance using different query fields with single word terms(JE-1)

The same runs as short query are examined:

1. Single word and phrasal terms with down-weighting for phrases and automatic feedback(JSCB2)

2. Single word and phrasal terms with down-weighting for phrases and no automatic feedback

3. Single word terms with automatic feedback(JSCB2-WD)

4. Single word terms with no automatic feedback

5. Single word and phrasal terms with normal tf*idf weighting and automatic feedback

6. Single word and phrasal terms with normal tf*idf weighting and no automatic feedback

The initial queries contain 42.2 single word terms and 12.5 phrasal terms in average and the final queries contain 73.3 single word terms and 35.0 phrasal terms in average.

Table 3 shows the results. Still the phrasal term with down-weighting run (JSCB2) outperformed phrasal term with normal tf*idf run, but the difference is smaller (3.5%-4.4% improvement).

Since initial queries are long enough and they contain rich terminology, performance improvements given by automatic feedback are also much smaller (o.4%-1.2%) than in short

Topic fields used	AFB	Avg. Prec	R-Prec
<title>,<description>,<narrative>,<concepts>	Yes	4249	4104
<title>,<description>,<narrative>,<concepts>	No	4230	4082
<title>,<description>,<narrative>	Yes	3798	3887
<title>,<description>,<narrative>	No	3691	3784
<title>,<description>,<concepts>	Yes	4068	4012
<title>,<description>,<concepts>	No	3956	3945
<title> <description>	Yes	3227	3267
<title> <description>	No	3082	3207
Phrasal terms with down-weighting (JSCB2)	Yes	4436	4301
Phrasal terms with down-weighting (JSCB1-normal TF*IDF)	Yes	3048 (3077)	3031 (3004)
Phrasal terms with down-weighting	No	4380	4309
<description>	No	2893	2914
Word terms(JSCB2-WD)	Yes	4184	4235
<title>	Yes	2538	2578
Word terms	No	4166	4227
<title>	No	2279	2448
Phrasal terms with normal tf*idf	Yes	4249	4104
Phrasal terms with normal tf*idf	No	4230	4082

Table 6: Performance using different query fields with phrasal terms and normal tf*idf(JE-1)

Table 3: Performance in Automatic Long Experiments (JE-1)

[1] These runs using only the <description> field are not identical to JSCB1 and its variations because they use the JSCB2 parameter set for the reason of comparison instead of the short query oriented parameter set of JSCB1.

query experiments (11%-13%).

Single word term run did not perform well this time.

It is interesting to see which topic field contributes to the performance of automatic long query retrieval.

As Table 4, Table 5 and Table 6 show, using only <title> and/or <description> fields, it is not clear if phrasal term runs are effective irrespective of with/without automatic feedback. Using <narrative> and/or <concepts> fields on top of them, phrasal term runs clearly outperform single word term runs.

We can see a correlation between the number of terms in original query (length of topic description) and the improvement from using phrasal terms.

Phrasal term indexing is effective only with enough long initial topic description (or enough rich terminology) containing a certain number of phrases as well as single words, otherwise its effect is rather incidental.

As phrasal terms have normally low frequency, short queries, that normally contain only one or two phrasal terms, their effects are probably less than the bias they may cause, consequently their effects are inconsistent.

4.3 Manual Query Run

The third official run that we named JSCB3, uses relevance feedback with automatic initial query construction, by which we intend to evaluate effects of relevance feedback in comparison with automatic run (JSCB2) performance.

The run was designed for simulating interactive retrieval processing where user interaction is restricted only for finding relevant documents from the relevance ranked list of the initial retrieval.

The ranked result lists of the first runs are examined by a searcher who finds relevant documents as many as possible in the time limit of 20 minutes for each query in view of relevance feedback. The time limit of 20 minutes, that we considered as reasonable and acceptable in real retrieval situation, was decided after the trial experiments by pre-test topics, in which we observe that 10 minutes are too short for a layman of the subject domain to understand the topic, to read each document carefully and to find as many as 10 documents per topic if possible; but 30 minutes are too long for non professional searchers to keep their concentration.

53 topics were divided into three sets and assigned for three people including the author of the article, a system engineer and a programmer. Searchers spent 17 minutes 26 seconds per topic in average looking for relevant (A-judgement) documents from the first retrieval result (using long query without automatic feedback), and found 539 possible relevant documents (10.17docs/topic).

Once such relevant document list, that is utilized only for term extraction purpose, is created, the system automatically constructs the final query for each topic processing relevance feedback from the relevant documents listed. Thus the final runs are executed in batch mode.

The comparison between our relevant document list and official relevance assessment files provided by the NTCIR organizer illustrates the reliability of our judgement as shown in Table 7.

If considered as a human aided retrieval system, our human judgement achieves 64% of precision while recall is 18% against JE-1 judgement assessment file. This performance is fitting on the precision-recall curves of roughly 40% average precision in our experiments data.

Again, the same runs as short query are examined:

1. Single word and phrasal terms with down-weighting for phrases and automatic feedback(JSCB3)

2. Single word and phrasal terms with down-weighting for phrases and no automatic feedback

3. Single word terms with automatic feedback(JSCB3-WD)

4. Single word terms with no automatic feedback

5. Single word and phrasal terms with normal tf*idf weighting and automatic feedback

6. Single word and phrasal terms with normal tf*idf weighting and no automatic feedback

The initial queries contain 42.2 single word terms and 12.5 phrasal terms in average exactly same as JSCB2 and after having expanded by the relevance feedback procedure, the final queries contain 58.8 single word terms and 42.2 phrasal terms in average.

The performance difference between different index language and weighting is relatively smaller in the final run after the relevance feedback than the initial retrieval as shown in Table 8.

As we have seen in long query experiments, when initial

Total	Judgement by NTCIR	Number of docs
539	A	344
	B	66
	C	129
	Unjudged	0

Table 7: Comparison between JSCB Team User Relevance Judgement and NTCIR Official Relevance Assessment

Run description	RFB	Avg. Prec	R-Prec
Phrasal terms with down-weighting (JSCB3)	Yes	4855	4648
Phrasal terms with down-weighting	No	4380	4309
Word terms(JSCB3-WD)	Yes	4776	4563
Word terms	No	4166	4227
Phrasal terms with normal tf*idf	Yes	4849	4753
Phrasal terms with normal tf*idf	No	4230	4082

Table 8: Performance in Manual Experiments(JE-1)

Topic fields used	RFB	Avg. Prec	R-Prec
<title>,<description>,<narrative>,<concepts> (=JSCB3)	Yes	4855	4648
<title>,<description>,<narrative>,<concepts>	No	4380	4309
<title>,<description>,<narrative>	Yes	4803	4672
<title>,<description>,<narrative>	No	3974	4065
<title>,<description>,<concepts>	Yes	4837	4750
<title>,<description>,<concepts>	No	4157	4084
<title>,<description>	Yes	4754	4716
<title>,<description>	No	3393	3621
<description>	Yes	4767	4704
<description>	No	3305	3442
<title>	Yes	4634	4521
<title>	No	2475	2735

Table 9: Performance using different query fields as initial query in manual runs (JE-1)

query is long and rich enough in terminology, the improvement given by the automatic feedback is limited although it never hurts the performance.

On the other hand, relevance feedback can gain 11% to 15% of improvements even when initial retrieval results are already at good level.

Table 9 shows that relevance feedback can increase avg. precision as high as 46%-48% even the initial query is as short as only “title” or “description” fields.

Despite the relatively low reliability (64% precision) of our human judgement, this relevance feedback procedure seems working well.

Our analysis for this fact is that even they are judged non relevant by NTCIR judges who are much more severe than our users for judgement, documents considered as relevant by a human user are similar to relevant documents and possibly share the same terminology with relevant documents, consequently they are useful for term extraction purpose.

Should this be true, we can assume that the relevance judgement for relevance feedback purpose is not necessarily very severe and some errors are totally allowable while such severe judgement is not realistic in end-user situation.

5. CONCLUSIONS

JSCB NTCIR experiments are described.

The following conclusions are drawn from these experiments:

1) Phrasal indexing is more effective when the initial topic description is long and rich in terminology.

2) Down-weighting for phrasal terms always merits the performance and it never hurts the performance in our experiments.

3) Automatic feedback also contributes for the performance especially when initial queries are short.

4) Relevance feedback technique adopted in JSCB3 is applicable for real retrieval situation and effective for high precision retrieval.

On the other hand, we need more experiments as well as careful observation on the effect of phrasal indexing with short queries.

For the future work, it is desirable to evaluate other weighting schemes instead of the empirical approach for down-weighting of phrasal terms adopted here.

6. ACKNOWLEDGMENTS

Our thanks to Mr. Toshiya Ueda and Mr. Tatsuo Kato for their assistance.

7. REFERENCES

[1] Cleverdon, C., Mills, J. and Keen, M. Factors Determining the Performance of Indexing Systems, vol. 1, Design Parts 1 and 2, Technical Report, Cranfield Institute of Technology; College of Aeronautics, Cranfield, 1966.

[2] Evans, D.A. and Lefferts, R.G., Grefenstette, G., Handerson, S.K., Hersh, W.R., and Archbold, A.A., CLARIT TREC Design, Experiments and Results, in Proceedings of the First Text REtrieval Conference(TREC-1), NIST Special Publication 500-207, Washington D.C., 1993, 494-501.

[3] Fagan, J.L. Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-syntactic Methods, Ph.D Thesis, Dept. of Computer Science, Cornell University, Sept. 1987.

[4] Lewis, D. Representation and Learning in Information Retrieval, Ph.D Thesis, Dept. of Computer and Information Science, University of Massachusetts, Feb. 1992.

[5] Lewis, D. An Evaluation of Phrasal and Clustered representation on a Text Categorization Task, in Proceedings of the Fifteenth Annual International ACM SIGIR Conference(Copenhagen, June 1992), ACM Press, 37-50.

[6] Milic-Frayling, N., Zhai, C., Tong, X., Jansen, P. and Evans, D.A. Experiments in Query Optimization: the CLARIT System TREC-6 Report, in Proceedings of the Sixth Text REtrieval Conference(TREC-6), NIST Special Publication 500-240, Washington D.C., 1998, 415-454.

[7] Salton, G. Automatic Text Processing, Addison-Wesley publishing company, Massachusetts ,1988.

[8] Strzalkowski, T., Carballo, J.P. and Marinescu, M. Natural Language Information Retrieval: TREC-3 Report, in Proceedings of the Third Text REtrieval Conference(TREC-3), NIST Special Publication 500-225, Washington D.C., 1995, 39-53

[9] Strzalkowski, T. and Sparck Jones, K. NLP Track at TREC-5, in Proceedings of the Fifth Text REtrieval Conference(TREC-5), NIST Special Publication 500-238, Washington D.C., 1997, 97-102.

[10] Strzalkowski, T., Guthrie, L., Karlgren, J., Leistensnider, J., Lin, F., Perez-Carballo, J., Straszheim, T., Wang, J. and Wilding, J. Natural Language Information Retrieval: TREC-5 Report, in Proceedings of the Fifth Text REtrieval Conference(TREC-5), NIST Special Publication 500-238, Washington D.C., 1997, 291-314

[11] Zhai, C., Tong, X., Milic-Frayling, N., Evans, D.A. Evaluation of Syntactic Phrase Indexing--CLARIT NLP Track Report, in Proceedings of the Fifth Text REtrieval Conference(TREC-5), NIST Special Publication 500-238, Washington D.C., 1997, 347-358

[12] Zhai, C. Fast Statistical Parsing of Noun Phrases for Document Indexing, in Proceedings of the 5th Conference on Applied Natural Language Processing, Washington D.C., 1997、312-319.

Appendix A. : Recall-Precision Graph of official submission runs and other experimental runs (JE-1, Interpolated 11 point precision)