The Effect of Cross-Lingual Pooling on Evaluation

Kazuko KURIYAMA Masaharu YOSHIOKA Noriko KANDO
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{kuriyama, yoshioka, kando}@nii.ac.jp

Abstract

The purpose of this study is to examine whether there is an effect on the relative evaluation of the IR systems using the relevance judgments made by the pooling method and additional interactive searches.

Relevance judgments of NTCIR-1&2 were made using the following steps: (1) collecting candidates for relevant documents by using the pooling method, (2) judging candidate documents by human assessors, (3)collecting additional candidates by recall-oriented interactive searches for search topics with more than 100 relevant documents to improve the exhaustiveness of the relevance judgments, and (4)judging the additional candidates.

For the purpose of the study we carried out experiments using the relevance judgments and search results submitted for the test of the 2nd NTCIR Workshop. First, we evaluated the search results using the final relevance judgments F of NTCIR-2 and $F - I$, that is, the F without the unique relevant documents found by the additional interactive searches I. Second, we made pools from the search results in each of the sub-tasks and evaluated the search results using the relevance judgments in the pools.

Almost the same rankings were produced by all the relevance judgments. Therefore our results verified the reliability of the evaluation using test collection based on pooling.

Keywords: *NTCIR-2, Pooling, Relevance Judgments, Reliability, Fairness*

1 Introduction

1.1 The Purpose of Our Experiments

For the construction of a large-scale test collection using the pooling method, there are many questions we must consider from the aspect of testing IR systems:

(1) exhaustiveness of the document pool,

(2) reliability of the test collection as a tool for system testing,

(3) inconsistency of relevance judgments.

In terms of (1) for the relevance judgments of NTCIR-1, pooling the top 100 documents from each search result worked well for topics with less than 50 relevant documents. For topics with more than 100 relevant documents, although the top 100 pooling covered only 51.9% of the total relevant documents for the pre-test, and 76.4% for the test of the 1st NTCIR Workshop, the coverage reached 89.7% and 98.0%, respectively, when combined with additional recall-oriented interactive searches [5],[6],[4].

In terms of (2) and (3), we found very high similarity among the system rankings produced using different sets of relevance judgments, regardless of the different coverage and pooling methods, and regardless of any inconsistency among relevance judgments [5],[6],[4].

In this paper, we examine whether there is an effect on the relative evaluation of the IR systems using the relevance judgments made by the pooling method and additional interactive searches.

For this purpose we carried out experiments using the relevance judgments and the search results submitted for the test of the 2nd NTCIR Workshop. First, we evaluated the search results using the final relevance judgments F of NTCIR-2 and F-I, that is, F without the unique relevant documents found by the additional interactive searches I. Second, we made pools from the search results for the sub-tasks and evaluated the search results using the relevance judgments in the pools.

1.2 Test Collections and Pooling Methods

A test collection for IR system testing consists of: (1) documents, (2) search topics, and (3) relevance judgments for each search topic. When constructing a test collection, it would be ideal to judge all documents for each search topic and make an exhaustive list of the relevant documents. However, this is not feasible for a large-scale database containing tens of thousands of documents.

The pooling method (Gilbert and Sparck Jones 1979[3]) is a well-known method for effectively and

efficiently collecting relevant document candidates for a large-scale test collection . In this approach, the top X documents retrieved by various systems using different retrieval algorithms for each topic are pooled, and then every document in a pool is judged by human assessors. Since 1992, the Text REtrieval Conference (TREC)[9][10],[12] has constructed large-scale test collections by the pooling method.

Recently, the Move-To-Front (MTF) pooling method was proposed as an improved variation of the pooling method (Cormack, Palmer and Clarke 1998[2]). Compared to the pooling method, the MTF pooling method prioritizes the search results and pools many more documents from the results with top priority, which are then judged. It has been shown that the MTF pooling method effectively produces a collection with considerably fewer judgments than would be required for the pooling method[2]. However, there remains a question for IR systems testing, as to whether it is unfair to change the number of documents pooled from each search result according to its priority.

Therefore we experimented with various pooling methods to verify the fairness of the test collection through the pooling method as a tool for testing IR systems.

2 Construction of Relevance Judgments of NTCIR-2

Relevance judgments of NTCIR-2 were made using the following steps: (1) collecting the candidates for relevant documents by using the pooling method, (2) judging candidate documents by human assessors, (3) collecting additional candidates by recall-oriented interactive searches for some search topics to improve the exhaustiveness of the relevance judgments, and (4) judging the additional candidates. Step (1) is too complicated to understand immediately. We show the steps and document collections used for the task in the following sub-sections.

2.1 Sub-tasks and Documents Used in Japanese & English IR Tasks

2.1.1 Sub-tasks

In the following, the "Japanese & English IR Task" is abbreviated to "JEIR Task". The search results were the outcomes of sub-tasks in two sub-categories of the JEIR Task. The sub-categories are Monolingual IR and Cross-Lingual IR.
The Monolingual IR includes:

- retrieval of Japanese documents by Japanese search topics (J-J Task)
- retrieval of English documents by English topics (E-E Task).

The Cross-Lingual IR includes:

- retrieval of Japanese documents by English topics (E-J Task)
- retrieval of English documents by Japanese topics (J-E Task)
- retrieval of a collection of a mixture of Japanese documents and English documents by either of Japanese topics (J-J,E Task) or English topics (E-J,E Task).

2.1.2 Document Collections Used for the Subtasks

Two documents collections, the J Collection and the E Collection were used for the JEIR Task in the 2nd NTCIR Workshop. The J Collection and the E Collection were extracted from two databases provided by the National Institute of Informatics (NII), *Academic Conference Papers Database* and *Grant-in-Aid Scientific Research Database*, a part of which are English-Japanese paired. The J Collection consists of three sets of documents, *ntc1-j1.mod*, *ntc2-j0g*, and *ntc2-j0k*. The E Collection consists of three sets of documents, *ntc1-e1.mod*, *ntc2-e0g*, and *ntc2-e0k*. The document sets *ntc1-j1.mod*, *ntc1-e1.mod*, *ntc2-j0g*, and *ntc2-e0g* were extracted from the *Academic Conference Papers Database*, and *ntc2-j0k* and *ntc2-e0k* were from the *Grant-in-Aid Scientific Research Database*.

When a Japanese document and an English document are paired in the original databases, they have the same document number "ACCN". In order to deal with J Collection and E Collection independently, we separated the paired documents into Japanese documents and English documents and attached new ACCNs to the English documents, that is, "gakkai-e-000040700" to "gakkai-e-000104007". The number of documents in the Document Collections and paired documents in J and E Collections are shown in Table 1.

Table 1. Number of documents in the Document Collections used for the 2nd NTCIR Workshop.

Document Collections	Number of docs
ntc1-j1.mod	332,918
ntc1-e1.mod	187,080
pairs in ntc1-j1&e1	181,485
ntc2-j0g	116,177
ntc2-e0g	77,433
pairs in ntc2-j0g&e0g	74,180
ntc2-j0k	287,071
ntc2-e0k	57,545
pairs in ntc2-j0k&e0k	57,512

Table 2. Relationship of Tasks, Search Topics, Documents, and Relevance Judgments.

Task	Topics	Document Collections	Relevance Judgments	
			Level1 (S or A)	Level2 (S, A or B)
J-J	topic-j101-150	ntc1-j1.mod, ntc2-j0g, ntc2-j0k	rel-j1.txt	rel-j2.txt
E-J	topic-e101-150			
J-E	topic-j101-150	ntc1-e1.mod, ntc2-e0g, ntc2-e0k	rel-e1.txt	rel-e2.txt
E-E	topic-e101-150			
J-J,E	topic-j101-150	ntc1-j1.mod, ntc2-j0g, ntc2-j0k, ntc1-e1.mod, ntc2-e0g, ntc2-e0k	rel-je1.txt	rel-je2.txt
E-J,E	topic-e101-150			

topic-j101-150 is the list of Japanese search topics. *topic-e101-150* is the list of English search topics.

The relationship of tasks, search topics, documents, and relevant judgments are shown in Table 2.

2.2 Pooling from the Runs

We refer to a search result as a *run* in the following sections.

We show a process for our pooling in Figure 1.

The participants of the sub-tasks retrieved documents from the J Collection and/or the E Collection for each search topic by their own IR systems and submitted the search results, that is, the runs. First we pooled the top X documents from the runs of all the sub-tasks, i.e., *cross-task* and *cross-lingual*, to collect candidates for relevant documents. The ACCNs of the Japanese and English documents in the pool were transformed to the original ACCNs, for example, "gakkai-j-0000407000" and "gakkai-e-000104007" to "gakkai-000040700"; that is, the English documents were mapped to the corresponding Japanese documents.

We see from Table 2 that both the J Collection and E Collection were used for the J-J,E task and the E-J,E task. If we pooled the top X documents from each run for all the tasks, a part of the collection of documents from the runs for the J-J,E task and the E-J,E task would overlap in the original databases and the total number of documents from the runs for the J-J,E task and the E-J,E task would, in reality, decrease. For an efficient pooling, we did not pool the documents from both the J-J,E task and E-J,E task except for one run each, which is the only one retrieved by a system of each participant for all the tasks.

Moreover, we selected two runs per participant according to their given priorities for each task. The reason we did not use all runs per participant is that empirically we think a system collects similar documents in its different runs and there are too many overlaps.

We show the number of the runs used for our real pooling in Table 3.

The number of documents pooled from each run X depends on the size of pool for each search topic; that is, X was adjusted from 70 to 100 so that the total number of documents for each topic might be less than 2000 to 2500. We show the Xs and the total number of documents in the pool P in Table 4. A pool $J1$ consists of Japanese documents pooled from the runs in J-J and E-J task. A pool $E1$ consists of English documents pooled from the runs in J-E and E-E task. A pool $J2$ consists of $J1$ and Japanese documents paired with documents in $E1$. A pool $E2$ consists of English documents paired with documents in $J1$ and $E1$. P is a pool in which the documents with original ACCNs were transformed from the documents in $J2$ and $E2$. $ave\%F$ is average coverage of the relevance documents in each pool.

Table 3. Number of submitted runs and pooled runs.

Task	Submitted runs	Pooled runs
J-J	93	29
J-E	41(1)	23(1)
J-J,E	15(1)	1
E-E	18	12
E-J	30	17
E-J,E	11	0

The "(n)"s are the numbers of the runs submitted by an organizer of the JEIR Task. Total numbers of submitted and pooled runs for the J-E task and J-J,E task include the "n"s.

The average coverage of the relevance documents in pools $J1$, $E1$, $J2$, and $E2$ shows that cross-lingual pooling is effective in collecting new relevant documents in paired languages. We mean that the cross-lingual pooling is (1) pooling documents from the runs of the task using one language document collection; that is, one of the J Collection and the E Collection, (2) transforming the ACCNs of Japanese documents to the ones of English documents and the ACCNs of the English documents to those of the Japanese documents, and (3) adding documents with new ACCNs to the pools $J1$ and $E1$ to get $J2$ and $E2$. (The definition of "relevant document" in this paper is detailed in the next sub-section.)

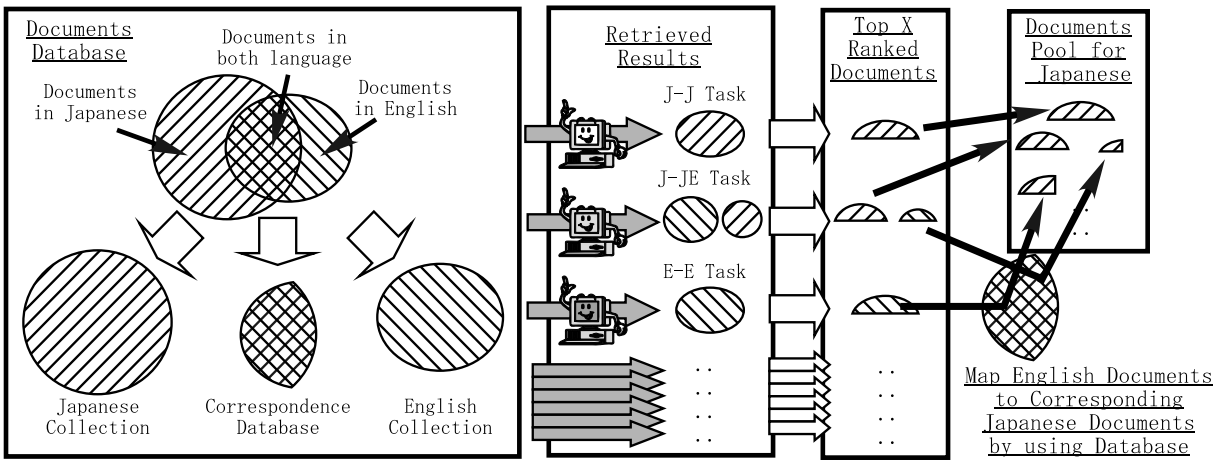

Figure 1. Process of Pooling.

2.3 Relevance Assessments

The relevance assessment for each topic was undertaken separately by two assessors, then cross-checked. The final judgments were based on negotiations between the two assessors and determined by the primary assessor of the topic (one of the two assessors), who created the topic. The judgments assign one of four grades, i.e., highly-relevant (S), relevant (A), partially-relevant (B), and non-relevant (C).

When the search of the topics had been made, five to 10 candidates of relevant documents for each topic were listed by the primary assessor who had created the topic.

The documents in pool P from the runs, and the unique results by the preliminary searches PP, which are not included in P, were judged separately by two assessors for each topic, and then cross-checked for 29 topics. Then final judgments for the pool P, and PP were determined by the primary assessors.

In the 2nd NTCIR Workshop we evaluated the runs by using the TREC's evaluation program. It was run against two different lists of relevant documents produced by two different thresholds, i.e., Level1, in which "S" and "A" are rated as "relevant", and Level2, in which "S", "A", and "B" are rated as "relevant". In the following we refer to the documents with a judgment "S", "A", or "B" as the "relevant documents".

2.4 Additional Interactive Searches

Additional recall-oriented interactive searches were carried out manually by graduate students who had majored in library and information science, for 16 topics with more than 110 relevant documents and/or the top 70 documents pooled from each run. Then the unique documents set I in the additional search results were judged by the primary assessors and added to the relevance judgments for the pool $P+PP$ to obtain the list of the final relevance judgments F.

We show the number of relevant documents in the pool P, the unique documents in the preliminary searches PP, the interactive searches I, and the final relevance judgments F for each search topic and average coverage of the relevance documents to F in Table 5. $ave\%all$ is average coverage of the relevance documents in each pool to F. $ave\%16$ is average coverage of the relevance documents in each pool to F for 16 search topics for which the additional interactive searches were carried out.

We see from Table 5 that the average coverage of J(P) and E(P) for the topics with more than 110 relevant documents, $ave\%16$ are 91.4% and 95.3%, respectively, and that is considered acceptable. This is due to there being many more runs submitted from the runs than for the pre-test and for the test of the 1st NTCIR-1. However, the recall-oriented interactive searches found 8.4% of the Japanese relevant documents J(F) and are effective to a degree.

3 Experimental Evaluation

3.1 System Testing Using Relevance Judgments with/without Additional Interactive Searches

To investigate whether the additional interactive searches have any effect on the system testing, we evaluated the runs for the J-J task and E-E task using the final relevance judgments F and $F-I$, which is the F without the unique relevant documents in the interactive search results I. We suppose that the additional interactive searches work as the runs from an IR system. To examine whether there are some effects on the evaluation by using the relevance judgments in the pool without the runs performed by a system, we

made two different pools $F-CRL$ and $F-DOVE$. The $F-CRL$ is the pool F without the unique documents in the runs submitted by a group *CRL*, and the $F-DOVE$ is the pool F without the unique documents in the runs submitted by a group *DOVE*. The reasons we selected the two groups are firstly that the CRL submitted the most runs, that is 27 runs, for the J-J task and all its runs used the automatic method for query construction; secondly, that the DOVE submitted many runs, that is 10 runs, for the J-J task and used automatic or interactive method for query construction; and thirdly, that they are suited for our purpose of investigating how the additional searches affect the evaluation of the runs submitted by automatic and interactive systems. The average coverage of the relevance documents in all the runs submitted by the CRL are 64.3% in Japanese relevance judgments *rel-j2.txt*, and 73.1% in English relevance judgments *rel-e2.txt*. The average coverage of the relevance documents in all the runs submitted by the DOVE are 47.7% in Japanese relevance judgments *rel-j2.txt* and 43.1% in English relevance judgments *rel-e2.txt*.

We show mean average precision and rankings of the runs for the J-J task and the E-E task produced by using F, $F-I$, $F-CRL$, and $F-DOVE$ in Table 6 and Table 7. Also we show the graphs of the mean average precisions in Figure 2 and Figure 3. Each run in the table is the run using the query field "DESCRIPTION" in the search topics, submitted by each participant for each task; that is, one top-ranked run using "D"per one system for each task was used for the evaluation, except for runs produced by the system which used the automatic method and the interactive method, respectively, for query construction, and when the system does not use the field "D".

We see in Table 6 and Table 7 that the same rankings are produced using different relevance judgments F, $F-I$, $F-CRL$ and $F-DOVE$ for the runs of the J-J task and the E-E task. In particular, each of four runs that used the interactive method has the same ranking over F and $F-I$. Therefore it is concluded that the additional interactive searches do not affect the system testing, regardless of their contributions to the exhaustiveness of the document pool for some particular topics.

3.2 System Testing Using Relevance Judgments in Pools from the Runs for Rach Task

To examine how the pools from the runs for the subtasks contribute to exhaustiveness and affect the relative system testing, we make some different pools and evaluate the runs for the J-J task and the E-E task.

3.2.1 Pooling

We extract documents in each part of the pool P for each sub-task, J-J, E-J, E-E, and E-J to make a pool for only one task. Numbers of the relevant documents in the pools P, $P(J-J)$ for only J-J task, $P(J-E)$ for only J-E task, $P(E-E)$ for only E-E task, and $P(E-J)$ for only E-J task are shown in Table 8. Also we make pools from F without the unique documents in the pools, $P(J-J)$, $P(J-E)$, $P(E-E)$, and $P(E-J)$, respectively; that is, $F-P(J-J)$, $F-P(J-E)$, $F-P(E-E)$, and $F-P(E-J)$. The numbers of the relevant documents in the pools are shown in Table 9.

We count the unique relevant documents from each system of the participants in the pools for the sub-tasks and all. The number of documents are shown in Table 10. The descriptor *all* means all four sub-tasks, J-J, J-E, E-E and E-J tasks, in the table. All systems in Table 10 found the relevant documents and contributed to the pools F.

3.2.2 System Testing

To investigate how the runs for the sub-task affect system testing, we evaluated the runs for the J-J task and E-E task by using the final relevance judgments F, the pools for the sub-tasks, $P(J-J)$, $P(J-E)$, $P(E-E)$, $P(J-E)$, and the pools, $F-P(J-J)$, $F-P(J-E)$, $F-P(E-E)$, and $F-P(E-J)$, which are the F without the unique relevant documents in each pool for each sub-task.

We show the mean average precisions and rankings of the runs for the J-J task and the E-E task produced by using F and the pools for the sub-tasks in Table 6 and Table 7. Also we show the graphs of the mean average precisions in Figure 2 and Figure 3.

We see in Table 6 and Table 7 that almost the same rankings are produced using different relevance judgments F and the other pools for the runs of the J-J task and the E-E task. Therefore it is concluded that the cross-lingual pooling and the unique contribution of the runs have little effect on the system testing, regardless of their contribution to the exhaustiveness of the document pool.

4 Summary and Conclusion

To investigate the fairness of the relevance judgements by the cross-lingual pooling and the additional recall-oriented interactive searches, we carried out experimental pooling and evaluations on the runs for the test of the 2nd NTCIR Workshop. From these experiments, our conclusions relating to the construction of NTCIR-2 are as follows.

(1) How the additional interactive searches affect the system testing: We scored the runs and ranked

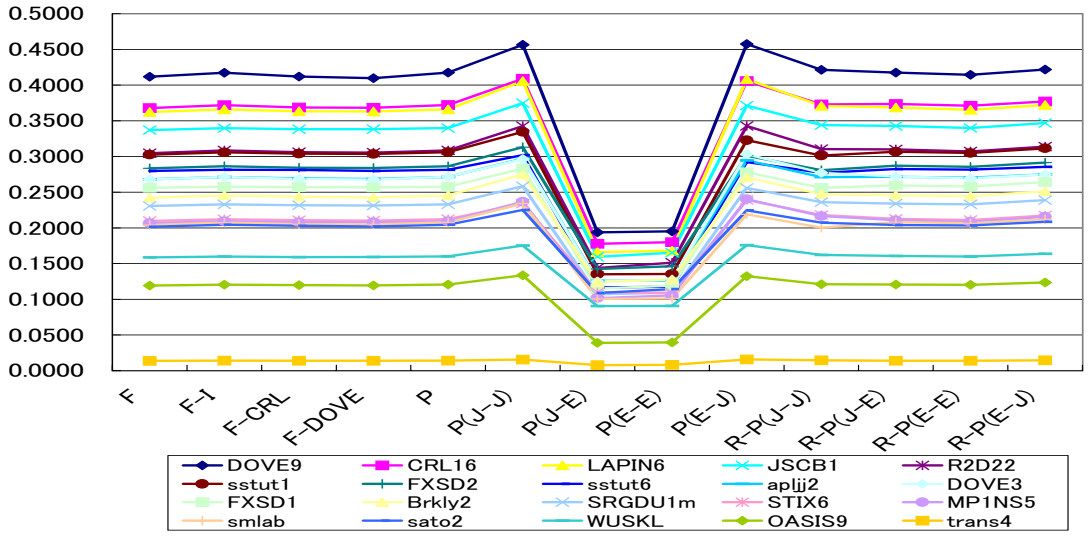

Figure 2. Graph of mean average precisions for the J-J task.

them by using the final relevance judgments F and $F-I$, which is the the relevance judgments without the unique documents I retrieved by the interactive searches for 16 search topics with more than 110 relevant documents and/or the top 70 documents pooled from each run. The rankings of the runs produced by their mean average precisions for the F and the $F-I$ are the same. This experimental result re-enforced our supossition that the interactive searches have no effect on the system testing.

(2) Whether the additional interactive searches are necessary or not: The average coverage of I for the 16 topics was 8.4% in the relevance judgments for the J Collection, *rel-j2.txt*. On the other hand, the average coverage of P, which is the pool from the runs, for all the topics reached 96.6% and 98.1%, respectively, for the J Collection and E Collection, that is, *rel-j2.txt* and *rel-e2.txt*. The interactive searches are effective judged against the exhaustiveness of the relevance judgments to a degree, but might not be necessary when the number of the pool runs is sufficiently large and diverse. However, we do not know the number of the pooled runs that is sufficient, and it is necessary to investigate this and the required variety.

(3) How the cross-lingual pooling affects the system testing and others: We scored the runs and ranked them by using the final relevance judgments F; the pools $P(J-J)$, $P(J-E)$, $P(E-E)$, and $P(E-J)$, which are the pools collected separately the documents from the runs for each sub-task. The rankings of the runs were produced by their mean average precisions for the relevance judgments in the pools.

For the J-E task and E-E task, the runs retrieved English documents which are in the E Collection, and we could obtain corresponding Japanese documents by mapping the English ones to the Japanese ones in J Collection. The corresponding Japanese documents are smaller sub-sets of all the relevance judgments than the J-J task and E-J task. Since the exhaustiveness of pools for the J-E task and the E-E task are lower than the J-J task and the E-J task, the average precision of the runs for the J-J task produced by the pools $P(J-E)$ and $P(E-E)$, are lower than those of the $P(J-J)$ and the $P(E-J)$. Although the contributions of the pools for the sub-task are different, the rankings produced by using the relevance judgments in the pools $P(J-J)$, $P(J-E)$, $P(E-E)$, and $P(E-J)$ are almost the same, regardless of the absolute magnitudes of the average precision of each evaluated run being different.

To examine how each pool for the sub-task has some effect on the evaluation and exhaustiveness, we made pools without the unique documents for each sub-task, $F-P(J-J)$, $F-P(J-E)$, $F-P(E-E)$, and $F-P(E-J)$. In the case of the $F-P(J-J)$, since the documents from the runs for the J-E, E-E, and E-J tasks make up for the loss of documents from the runs for the J-J task, it has no effect on the system evaluation; similarly for the cases of the $F-P(J-E)$, $F-P(E-E)$ and $F-P(E-J)$. Hence we can say that the loss of the documents from almost all the runs for the

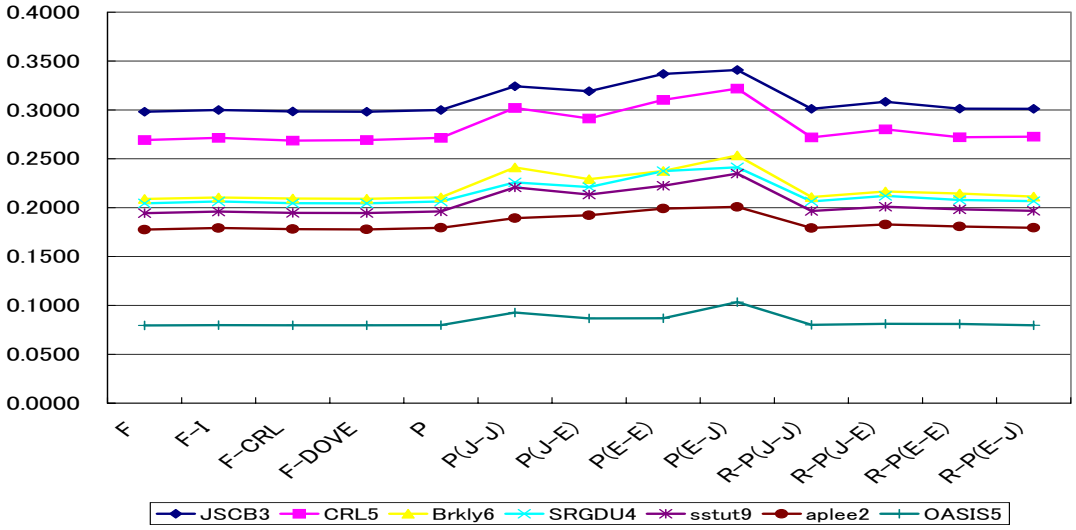

Figure 3. Graph of mean average precisions for the E-E task.

J-J,E task and the E-J,E task might have no effect on the exhaustiveness of the pool and the evaluation. There is a unique contribution of the runs by a system (participant) to a degree. The contribution does not always improve the rankings of the runs, but it seems to affect the rankings.

Finally, we conclude that the cross-lingual pooling, which is pooling mono-lingual documents and mapping them to corresponding documents in another language, has little effect on the system evaluation, and it is useful to collect candidates of relevant documents effectively by using the correspondence.

In conclusion, we should note that the average precision (and ranking produced by it) as a measure for evaluating retrieval performance is very robust and we should try another measure for evaluation.

5 Acknowledgment

We thank Prof. Kazuaki Kishida for his substantial advice.

This research is a part of the research project “A Study on Ubiquitous Information System for Utilization of Highly Distributed Information Resources”, supported by JSPS (Japan Society for the Promotion of Science) grant, JSPS-RFTF96P00602.

References

[1] Buckley, C., Voorhees, E., “Tutorial: Theory and Practice in Text Retrieval System Evaluation”. ACM–SIGIR’99, Berkeley, CA U.S.A, 1999.

[2] Cormack, G.V. et al., “Efficient Construction of Large Test Collections”. In Proc. ACM–SIGIR’98, pp.282–289, Melbourne, 1998.

[3] Gilbert, G., Sparck Jones, K., “Statistical Bases of Relevance Assessment for the ‘Ideal’ Information Retrieval Test Collection”. BL R&D Report 5481, 1979.

[4] Kando, N., Kuriyama, K., Nozue, T., “NTCIR-1 (NACSIS Test Collection for Information Retrieval systems-1): Its policy and practice”. IPSJ SIG Notes, No.99–FI–53–5, pp.33–40, 1999. (In Japanese.)

[5] Kuriyama, K. et al., “Pooling for a Large Scale Test Collection: Analysis of the Search Results for the Pre-test of the NTCIR-1 Workshop”. IPSJ SIG Notes, No.99–FI–54–4, pp.25–32, 1999. (In Japanese.)

[6] Kuriyama, K. et al., “NACSIS Test Collection for Information Retrieval systems-1 (1): Analysis of the Pooling and the Relevance Assessments”. In Proc. IPSJ Annual Meeting, Vol.3, pp.105-106, 1999. (In Japanese.)

[7] NII-NACSIS Test Collection for Information Retrieval systems. http://research.nii.ac.jp/ntcir/

[8] NTCIR Workshop 1: Proceedings of the First NTCIR Workshop on Retrieval in Japanese Text Retrieval and Term Recognition, Tokyo, Japan, Aug.30–Sep.1, 1999, ISBN 4-924600-77-6.

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings/

[9] Text REtrieval Conference (TREC). http://trec.nist.gov/ (visited January 12th, 2001).

[10] Voorhees, E., "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness". In Proc. ACM–SIGIR'98, pp.315–232, Melbourne, 1998.

[11] Voorhees, E., Harman, D., "Overview of the Eighth Text REtrieval Conference (TREC-8)". NIST Special Publication 500-246.

[12] Voorhees, E., Harman, D. eds. The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-242, Maryland, U.S.A., 2000.

[13] Zobel, J., "How Reliable are the Results of Large Scale Information Retrieval Experiments?". In Proc. ACM–SIGIR'98, pp.307–314, Melbourne, 1998.

Table 4. Number of documents in pool P.

topic	X	$J1$	$E1$	$J2$	$E2$	P
0101	100	1185	899	1889	1091	1905
0102	100	1137	1024	1963	1181	1997
0103	90	1361	846	2019	1069	2033
0104	100	1240	827	1883	978	1895
0105	80	1214	921	2030	1051	2050
0106	70	1199	990	2049	1181	2085
0107	100	1110	806	1739	963	1769
0108	100	1228	848	1751	1129	1770
0109	100	1091	1063	1897	1277	1932
0110	90	1458	986	2109	1434	2125
0111	70	1350	1140	2264	1569	2298
0112	80	1038	1099	2018	1248	2037
0113	100	1121	1109	2069	1376	2103
0114	90	1237	1070	2059	1354	2079
0115	70	1702	1305	2795	1936	2824
0116	100	1173	689	1600	1185	1602
0117	80	1307	976	2066	1249	2079
0118	100	1154	1069	1916	1277	1936
0119	70	1417	1470	2712	1907	2716
0120	80	1299	1086	2107	1431	2150
0121	100	1532	888	2081	1280	2113
0122	100	1255	952	1967	1169	1970
0123	70	1025	1026	1909	1263	1919
0124	100	1079	1056	1979	1179	2009
0125	70	1175	1113	2130	1410	2146
0126	80	1054	1258	2158	1504	2167
0127	100	1142	800	1761	957	1781
0128	100	1261	990	2107	1224	2124
0129	70	1170	1300	2297	1567	2328
0130	70	1508	1483	2755	1844	2803
0131	100	1079	966	1868	1112	1896
0132	80	1326	842	1976	1029	2014
0133	90	1050	1238	2121	1499	2133
0134	100	895	1126	1824	1308	1848
0135	80	1121	1193	2154	1483	2163
0136	100	1180	1115	2137	1356	2173
0137	70	1488	996	2266	1321	2292
0138	90	1221	1012	2051	1264	2085
0139	100	1379	911	2013	1238	2039
0140	80	1394	927	2127	1143	2148
0141	100	1354	946	2099	1228	2108
0142	70	1505	1229	2589	1625	2605
0143	100	1320	991	2022	1313	2036
0144	70	1392	827	2007	1343	2062
0145	100	882	777	1449	1061	1456
0146	70	1650	1257	2689	1708	2724
0147	70	1695	1086	2548	1701	2574
0148	70	1397	960	2047	1607	2063
0149	80	1324	881	2033	1176	2057
ave % F		89.6	91.8	96.6	98.8	100

X is the numbet of documents pooled from each run. A pool $J1$ consists of Japanese documents pooled from the runs in J-J and E-J task. A pool $E1$ consists of English documents pooled from the runs in J-E and E-E task. A pool $J2$ consists of $J1$ and Japanese documents paired with documents in $E1$. A pool $E2$ consists of English documents paired with documents in $J1$ and $E1$. P is a pool in which the documents with original ACCNs were transformed from the documents in $J2$ and $E2$. $ave\%F$ is average coverage of the relevance documents in each pool.

Table 5. Number of relevant documents in the pools.

topic	J(P)	J(PP)	J(I)	J(F)	E(P)	E(PP)	E(I)	E(F)
0101	96			96	24			24
0102	23			23	11			11
0103	26			26	12			12
0104	41			41	9			9
0105	41			41	6			6
0106	35			35	7			7
0107	44			44	9			9
0108	101			101	78			78
0109	48			48	19			19
0110	110		5	115	69		2	71
0111	140	1	90	231	99	1	25	125
0112	111		9	120	15		0	15
0113	118			118	12			12
0114	45			45	21			21
0115	183		26	209	119		11	130
0116	41			41	32			32
0117	123		3	126	37		0	37
0118	71			71	32			32
0119	168		10	178	68		2	70
0120	26			26	11			11
0121	202			202	105			105
0122	19			19	5			5
0123	61		0	61	17		0	17
0124	84	1		85	17			17
0125	30			30	11			11
0126	151		15	166	49		1	50
0127	144		23	167	38		2	40
0128	23			23	8			8
0129	57			57	16			16
0130	86			86	23			23
0131	152			152	20			20
0132	297	5	39	341	137	4	15	156
0133	160		2	162	33		0	33
0134	140			140	32			32
0135	196		15	211	65		0	65
0136	48			48	15			15
0137	15			15	7			7
0138	92			92	53			53
0139	225		0	225	182		0	182
0140	209		8	217	68		0	68
0141	204			204	58			58
0142	41			41	22			22
0143	24			24	11			11
0144	78		16	94	51		8	59
0145	22			22	19			19
0146	12			12	9			9
0147	265	1	101	367	196	1	45	242
0148	55			55	44			44
0149	68	1		69	25	2		27
ave % all	96.6	0.1	3.3	100	98.1	0.2	1.7	100
ave % 16	91.4	0.2	8.4	100	95.3	0.7	4.0	100

J($pool$) is the Japanese relevant documents in the pool $pool$. E($pool$) is the English relevant documents in the pool $pool$. The $pool$ is each of P, PP, I, and F. $ave\%all$ is average coverage of the relevance documents in each pool to F. $ave\%16$ is average coverage of the relevance documents in each pool to F for 16 search topics for which the additional interactive search was carried out.

Table 6. Mean average precisions and rankings of the runs for the J-J task.

Run-ID	DOVE9	CRL16	LAPIN6	JSCB1	R2D22	sstut1	FXSD2	sstut6	apljj2	DOVE3
Query Field	D N	D	D	D	D	D	T D N C F	D	D	D
Method	interact	auto	auto	auto	auto	auto	interact	interact	auto	auto
F	1	2	3	4	5	6	7	8	9	10
	0.4118	0.3679	0.3623	0.3370	0.3046	0.3026	0.2834	0.2797	0.2680	0.2678
F-I	1	2	3	4	5	6	7	8	9	10
	0.4173	0.3720	0.3659	0.3396	0.3085	0.3059	0.2863	0.2814	0.2713	0.2713
F-CRL	1	2	3	4	5	6	7	8	9	10
	0.4120	0.3686	0.3634	0.3382	0.3061	0.3039	0.2842	0.2811	0.2692	0.2685
F-DOVE	1	2	3	4	5	6	7	8	9	10
	0.4097	0.3682	0.3630	0.3382	0.3055	0.3035	0.2838	0.2798	0.2685	0.2677
P	1	2	3	4	5	6	7	8	10	9
	0.4175	0.3721	0.3660	0.3398	0.3086	0.3061	0.2863	0.2815	0.2714	0.2715
P(J-J)	1	2	3	4	5	6	7	8	9	10
	0.4565	0.4088	0.4056	0.3747	0.3425	0.3345	0.3130	0.3019	0.2970	0.2964
P(J-E)	1	2	3	4	5	7	6	11	8	9
	0.1938	0.1777	0.1661	0.1595	0.1439	0.1351	0.1425	0.1166	0.1258	0.1251
P(E-E)	1	2	3	4	5	7	6	12	10	9
	0.1951	0.1801	0.1682	0.1652	0.1512	0.1354	0.1464	0.1177	0.1261	0.1271
P(E-J)	1	3	2	4	5	6	8	10	9	7
	0.4575	0.4056	0.4082	0.3713	0.3428	0.3228	0.3011	0.2922	0.2945	0.3031
F-P(J-J)	1	2	3	4	5	6	7	8	10	9
	0.4213	0.3729	0.3709	0.3441	0.3104	0.3013	0.2807	0.2825	0.2715	0.2720
F-P(J-E)	1	2	3	4	5	6	7	8	10	9
	0.4175	0.3735	0.3686	0.3428	0.3100	0.3066	0.2871	0.2770	0.2711	0.2770
F-P(E-E)	1	2	3	4	5	6	7	8	9	10
	0.4145	0.3710	0.3654	0.3398	0.3071	0.3051	0.2856	0.2816	0.2704	0.2696
F-P(E-J)	1	2	3	4	5	6	7	8	9	10
	0.4218	0.3771	0.3722	0.3468	0.3139	0.3114	0.2915	0.2855	0.2750	0.2749

Run-ID	FXSD1	Brkly2	SRGDU1m	STIX6	MP1NS5	smlab	sato2	WUSKL	OASIS9	trans4
Query Field	D	D	D	D	D	D N C	D	D	D	D
Method	auto	auto	auto	auto	auto	interact	auto	auto	auto	auto
F	11	12	13	14	15	16	17	18	19	20
	0.2561	0.2421	0.2307	0.2097	0.2067	0.2040	0.2015	0.1587	0.1192	0.0138
F-I	11	12	13	14	15	16	17	18	19	20
	0.2579	0.2454	0.2329	0.2121	0.2093	0.2076	0.2044	0.1599	0.1205	0.0141
F-CRL	11	12	13	14	15	16	17	18	19	20
	0.2572	0.2437	0.2318	0.2108	0.2079	0.2052	0.2030	0.1591	0.1198	0.0140
F-DOVE	11	12	13	14	15	16	17	18	19	20
	0.2568	0.2427	0.2314	0.2102	0.2079	0.2051	0.2019	0.1593	0.1194	0.0139
P	11	12	13	14	15	16	17	18	19	20
	0.2579	0.2455	0.2330	0.2122	0.2093	0.2077	0.2044	0.1600	0.1206	0.0141
P(J-J)	11	12	13	16	14	15	17	18	19	20
	0.2831	0.2750	0.2584	0.2339	0.2367	0.2342	0.2251	0.1753	0.1337	0.0156
P(J-E)	12	10	15	13	16	17	14	18	19	20
	0.1152	0.1241	0.1070	0.1099	0.1017	0.1005	0.1088	0.0906	0.0390	0.0078
P(E-E)	11	8	14	15	16	17	13	18	19	20
	0.1190	0.1278	0.1094	0.1088	0.1055	0.1012	0.1144	0.0908	0.0396	0.0082
P(E-J)	11	12	13	15	14	17	16	18	19	20
	0.2780	0.2707	0.2559	0.2392	0.2401	0.2190	0.2245	0.1759	0.1324	0.0158
F-P(J-J)	11	12	13	14	15	17	16	18	19	20
	0.2560	0.2472	0.2361	0.2180	0.2166	0.2003	0.2074	0.1622	0.1211	0.0146
F-P(J-E)	11	12	13	14	15	16	17	18	19	20
	0.2597	0.2457	0.2339	0.2126	0.2107	0.2082	0.2040	0.1608	0.1206	0.0140
F-P(E-E)	11	12	13	14	15	16	17	18	19	20
	0.2580	0.2444	0.2329	0.2111	0.2085	0.2064	0.2034	0.1599	0.1203	0.0139
F-P(E-J)	11	12	13	14	15	16	17	18	19	20
	0.2642	0.2513	0.2389	0.2173	0.2156	0.2120	0.2086	0.1637	0.1235	0.0145

Query Field shows the field(s) of the search topics used for the runs.

Table 7. Mean average precision and ranking of the runs for the E-E task.

Run-ID	JSCB3	CRL5	Brkly6	SRGDU4	sstut9	aplee2	OASIS5
Query Field	D	D	D	T D	D	D	D
Method	auto	auto	auto	auto	auto	auto	auto
F	1	2	3	4	5	6	7
	0.2981	0.2692	0.2089	0.2044	0.1944	0.1776	0.0795
F-I	1	2	3	4	5	6	7
	0.2999	0.2715	0.2104	0.2063	0.1961	0.1793	0.0798
F-CRL	1	2	3	4	5	6	7
	0.2984	0.2686	0.2094	0.2046	0.1947	0.1780	0.0797
F-P(E-J)	1	2	3	4	5	6	7
	0.2981	0.2692	0.2090	0.2044	0.1945	0.1777	0.0796
P	1	2	3	4	5	6	7
	0.3000	0.2715	0.2105	0.2064	0.1962	0.1794	0.0798
P(J-J)	1	2	3	4	5	6	7
	0.3242	0.3021	0.2411	0.2258	0.2207	0.1893	0.0927
P(J-E)	1	2	3	4	5	6	7
	0.3192	0.2912	0.2291	0.2211	0.2134	0.1923	0.0867
P(E-E)	1	2	3	4	5	6	7
	0.3368	0.3103	0.2376	0.2374	0.2224	0.1991	0.0868
P(E-J)	1	2	3	4	5	6	7
	0.3409	0.3218	0.2535	0.2413	0.2348	0.2008	0.1033
F-P(J-J)	1	2	3	4	5	6	7
	0.3012	0.2719	0.2109	0.2063	0.1967	0.1792	0.0801
F-P(J-E)	1	2	3	4	5	6	7
	0.3083	0.2801	0.2166	0.2121	0.2010	0.1828	0.0812
F-P(E-E)	1	2	3	4	5	6	7
	0.3013	0.2721	0.2144	0.2078	0.1983	0.1808	0.0810
F-P(E-J)	1	2	3	4	5	6	7
	0.3012	0.2726	0.2113	0.2067	0.1968	0.1794	0.0797

Query Field shows the field(s) of the search topics used for the runs.

Table 8. Number of relevant documents in the pools for sub-tasks.

topic	J($P(J-J)$)	J($P(J-E)$)	J($P(E-E)$)	J($P(E-J)$)	E($P(J-J)$)	E($P(J-E)$)	E($P(E-E)$)	E($P(E-J)$)
0101	78	23	23	88	20	23	23	21
0102	23	10	10	23	11	11	11	11
0103	23	12	12	24	10	12	12	11
0104	41	9	9	40	9	9	9	9
0105	39	6	6	33	4	6	6	3
0106	35	5	6	28	7	5	6	4
0107	42	8	8	31	8	8	8	9
0108	89	75	57	69	65	76	58	54
0109	48	15	15	47	16	18	18	16
0110	71	54	52	83	44	52	52	51
0111	74	73	58	61	44	74	58	33
0112	97	15	16	95	10	15	15	11
0113	110	15	14	102	10	12	11	9
0114	37	19	18	31	15	19	17	14
0115	145	66	38	51	84	65	37	30
0116	39	29	26	38	30	29	26	31
0117	104	33	28	79	30	33	28	23
0118	59	33	25	64	27	29	22	26
0119	140	36	27	85	53	36	27	31
0120	26	11	11	23	11	11	11	11
0121	161	90	77	151	72	92	80	66
0122	14	5	5	18	3	5	5	4
0123	60	14	10	47	16	16	10	12
0124	80	18	18	69	17	17	17	13
0125	30	8	5	26	8	11	5	7
0126	123	39	43	107	26	38	42	28
0127	133	33	31	118	33	33	31	30
0128	21	9	9	23	8	8	8	8
0129	54	16	18	41	14	14	16	12
0130	74	14	14	26	13	14	15	5
0131	144	26	26	133	18	20	20	18
0132	168	119	93	119	38	116	92	18
0133	148	45	44	132	29	33	32	30
0134	120	42	35	96	20	30	27	16
0135	142	56	62	123	33	52	59	29
0136	43	16	15	36	12	15	14	11
0137	14	5	4	11	4	6	5	4
0138	59	52	41	49	24	50	39	17
0139	183	155	129	189	145	154	128	152
0140	115	52	42	146	28	50	41	31
0141	132	50	22	136	31	47	23	28
0142	28	16	20	31	13	14	19	13
0143	22	9	8	21	9	9	8	10
0144	48	38	28	53	26	44	32	27
0145	22	20	20	22	19	19	19	19
0146	12	8	8	12	8	9	9	8
0147	152	145	137	148	102	148	139	98
0148	38	34	38	42	29	34	38	34
0149	43	25	12	31	10	22	9	7
ave % F	81.6	35.3	31.9	73.5	72.9	86.5	78.8	67.3

$ave\%F$ is average coverage of the relevance documents in each pool.

Table 9. Number of relevant documents in the pools without a sub-task.

topic	J($F-P(J-J)$)	J($F-P(J-E)$)	J($F-P(E-E)$)	J($F-P(E-J)$)	E($F-P(J-J)$)	E($F-P(J-E)$)	E($F-P(E-E)$)	E($F-P(E-J)$)
0101	91	96	96	82	24	24	24	24
0102	23	23	23	23	11	11	11	11
0103	25	26	26	25	12	12	12	12
0104	40	41	41	41	9	9	9	9
0105	36	41	41	41	6	6	6	6
0106	31	35	35	35	7	7	7	7
0107	31	44	44	43	9	9	9	9
0108	91	92	101	101	77	69	78	78
0109	47	48	48	48	19	19	19	19
0110	106	109	111	102	68	67	67	69
0111	209	213	223	218	116	106	117	123
0112	109	120	119	112	15	15	15	15
0113	106	117	118	114	12	11	12	12
0114	37	44	44	41	21	19	20	20
0115	131	192	203	204	94	112	124	128
0116	39	40	41	40	32	31	32	31
0117	97	122	125	114	37	33	36	36
0118	71	69	71	63	32	30	32	30
0119	121	172	175	160	59	64	67	65
0120	23	26	26	26	11	11	11	11
0121	183	194	199	191	100	98	102	104
0122	18	19	19	15	5	5	5	5
0123	50	61	61	60	17	16	17	17
0124	74	85	85	81	17	17	17	17
0125	27	30	30	30	11	8	11	11
0126	143	162	157	162	49	46	41	50
0127	148	166	167	160	40	39	40	40
0128	23	23	23	21	8	8	8	8
0129	45	57	56	56	16	16	15	16
0130	42	84	81	84	20	21	18	23
0132	281	310	328	313	151	126	144	155
0133	139	161	162	156	33	32	33	33
0134	117	134	140	138	30	29	32	32
0135	187	206	200	202	65	62	55	65
0136	41	47	48	47	15	14	15	15
0137	13	14	15	15	7	6	7	7
0138	86	81	89	89	53	43	50	53
0139	215	213	224	212	177	170	181	174
0140	190	202	211	161	66	53	62	65
0141	165	191	204	160	53	48	58	54
0142	39	41	36	38	21	22	17	21
0143	21	24	24	23	10	11	11	10
0144	88	88	93	83	58	51	58	57
0145	22	22	22	22	19	19	19	19
0146	12	12	12	12	9	9	9	9
0147	331	352	357	335	230	225	231	234
0148	51	55	52	51	43	44	41	43
0149	53	60	68	62	25	18	27	26
ave % F	87.8	97.3	98.6	94.4	97.4	93.3	96.9	98.5

$ave\%F$ is average coverage of the relevance documents in each pool.

Table 10. Number of unique relevant documents from the systems for the sub-tasks.

Group's ID	Part.Tasks	J($P(J-J)$)	J($P(J-E)$)	J($P(E-E)$)	J($P(E-J)$)	J(P)	E($P(J-E)$)	E($P(J-J)$)	E($P(E-E)$)	E($P(E-J)$)	E(P)
apl	all	51	50	93	147	98	48	38	92	59	55
ATT	J-E, E-J		7		30	12	5			13	3
Brkly	all	44	42	14	157	69	40	16	14	52	30
CAMUK	J-E		24			11	22				10
CRL	all	67	48	42	164	95	46	37	39	63	29
DLUT	J-E, E-J		11		21	12	11			5	6
DOVE	J-J	118				56		40			9
FXSD	J-J	68				45		16			6
Forst	J-E, E-J		39		132	77	38			76	39
JSCB	all	42	60	68	205	109	56	10	66	58	41
LAPIN	J-J	48				22		13			2
LISIF	J-E, E-J		17		46	16	15			13	4
MP1NS	J-J, J-E	61	80			57	78	13			34
NTHU	J-E		5			2	4				1
R2D2	J-J	30				11		17			4
SRGDU	J-J, E-E	37		45		38		10	42		20
sato	J-J	45				14		18			2
smlab	J-J	71				50		22			12
sstut	all	57	53	41	216	108	51	16	41	60	42
STIX	J-J	6				5		0			0
trans	J-J	30				18		12			5