大規模テストコレクション NTCIR-2 の構築 ——対話型追加検索と言語横断的プーリングの効果

栗山和子[†] 吉岡真治[††] 神門典子[†]

大規模テストコレクション NTCIR-2 の正解文書リストは，NTCIR ワークショップ 2 において各参加者から提出された検索結果を用いて，プーリング法に基づいて作成された．本稿では，NTCIR-2 の正解文書リストの作成過程において行われた，言語横断的プーリング，および対話型検索システムを用いた追加検索が，検索システムの NTCIR-2 を用いた相対的評価にどのような影響を与えるかを考察する．また，プーリングに参加しなかったシステムを評価する場合に，プーリング法で作成した NTCIR-2 が有効であるかどうか調べる．本研究では，NTCIR-2 の正解文書リストと，NTCIR ワークショップ 2 の参加チームの提出結果を用いて評価実験を行った．まず，最終的な正解文書リスト F と，F から追加の対話型検索 I だけで見つかった文書を除いた正解文書リスト $F-I$ を用いて，提出結果の評価を行った．次に，各サブタスクごとの提出結果からプーリングを行い，このサブタスクごとのプールを正解文書リストとして用いた場合の評価を行った．さらに，プーリングに参加しなかったシステムの評価をシミュレートするため，F から同じシステムの提出結果の集合だけに含まれている正解文書 S を除いた正解文書リスト $F-S$ を用いて，そのシステムの提出結果を評価した．いずれの場合でも，提出結果の平均精度の平均による順位付けを行い，相対的評価とした．結果として，どの文書リストを正解文書リストとして用いて提出結果の評価を行っても，提出結果の相対的な順位はほとんど変化しなかった．また，そのシステムだけが見つけた正解文書を除いても，すなわち，そのシステムがプーリングに参加しなくても，そのシステムの提出結果の評価にはほとんど影響がなく，プーリングに参加した他システムとの相対的評価についても影響がないことが分かった．このことから，プーリング法に基づいて作成したテストコレクションの信頼性を確かめることができた．

Construction of a Large Scale Test Collection NTCIR-2 ——The Effect of Additional Interactive Search and Cross-Lingual Pooling

Kazuko Kuriyama,[†] Masaharu Yoshioka[††] and Noriko Kando[†]

The purposes of this study are to examine whether there is an effect on the relative evaluation of the IR systems using the relevance judgments of the test collection NTCIR-2 made by the pooling method and additional interactive searches, and to investigate whether the NTCIR-2 is effective for evaluating the IR systems, the search results of which were not used for the pooling. We carried out experiments using different lists of relevance judgments and search results submitted for the test of the 2nd NTCIR Workshop. First, we evaluated the search results using the list of 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 for each of the sub-tasks and evaluated the search results using the pools as lists of relevance judgments. Third, we evaluated the search results using $F-S$, that is, the F without the unique relevant documents found by an IR system S in order to simulate evaluation of the IR system which was not used for the pooling. Almost the same rankings of the search results were produced by using the pools as lists of relevance judgments for system evaluation. When the search results by an IR system were not used for the pooling, there is a very little effect on evaluation of the system itself and relative evaluation among it and other systems. Therefore our results verified the reliability of test collection as an evaluation tool, which was based on pooling method.

1. はじめに

1.1 NTCIR プロジェクト

著者らは，国立情報学研究所（旧学術情報センター）「情報検索システム評価用テストコレクション構築プロ

† 国立情報学研究所
National Institute of Informatics (NII)
†† 北海道大学大学院工学研究科
Graduate School of Engineering, Hokkaido University

ジェクト」において，情報検索システム評価用テストコレクション NTCIR（エンティサイル：NII-NACSIS Test Collection for Information Retrieval systems）の構築を行っている[8]．その過程において，1998 年 11 月から 1999 年 9 月まで，テストコレクション 1（NTCIR-1）を用いた評価型ワークショップ，NTCIR ワークショップ 1[9] を開催し，2000 年 5 月から 2001 年 3 月まで，テストコレクション 2（NTCIR-2）を用いた，NTCIR ワークショップ 2[10] を開催し，テストコレクションの構築および検索システムの評価を行ってきた．

1.2 目　　的

テストコレクションとは，情報検索システムの検索性能評価に用いられる実験用セットのことであり，(1) 文書データベース，(2) 検索課題群，(3) 各検索課題に対する正解文書の網羅的リスト，からなる．

大規模テストコレクションの正解文書リストの構築法としては，各検索課題ごとに，複数の異なる検索結果の上位一定数の文書をプールし，それを人間の正解判定者が検索課題に適合（正解）か不適合（不正解）かを判定して，正解文書のリストを作成する，プーリング法が一般的である．

プーリング法による大規模テストコレクションの構築については，情報検索システムの評価という側面から以下のような点について考慮する必要がある．

(1) 正解文書リストの網羅性：
プーリングによる正解文書収集では，プールに入れられなかった文書は不正解文書であるものと仮定される．そのため，正解文書候補をいかに網羅的に集めてプールすることができるかということが問題となる．

(2) 正解文書リストの公平性：
検索システムの評価という観点から，正解文書リストはどのような検索システムに対しても公平になるような方法で作成する必要がある．

(3) 正解判定の無矛盾性：
正解判定が複数の判定者によって行われるとき，判定者間の判定にはゆれがある．判定者が異なる正解判定リストをそれぞれ使用したとき，その判定のゆれによって，システムの相対的評価がどのような影響を受けるかを検証する必要がある．

筆者らは，テストコレクション NTCIR-1 構築の過程において，上記の点について実験と考察を行った[4)~7)]．(1) について，NTCIR-1 では，上位一定数のプーリングと対話型検索システムを用いた追加検索によって正解文書リストの網羅性を高めることができた．具体的には，正解文書が 100 件以上の検索課題については，各提出結果からの上位 100 件ずつのプーリングでは，NTCIR ワークショップ 1[9] の予備テストでは全正解文書数の 51.9%，評価テストでは 76.4%しか網羅できなかったが，対話型検索システムで手作業で再現率を重視した追加検索を行った結果を追加すると，予備テストでは 89.7%，評価テストでは 98.0%をカバーすることができた．(2) に関しては，予備テストと評価テストの提出結果のそれぞれについて，プーリングと追加検索によって作成した数種類の正解文書リストを用いて評価を行い，プーリングと追加検索による正解文書リストの作成が相対的なシステム評価に影響を与えないことを確かめた．また，(3) については，異なる 2 人の正解判定者による判定結果と最終判定結果という 3 つの異なる正解判定結果リストを用いて評価テストの提出結果を評価し，その結果，異なる判定結果による相対的なシステム評価（各提出結果の順位）は変わらないので，異なる判定者間の判定の違いは，システム評価にはほとんど影響を与えないということが確かめられた．

本稿では，NTCIR-1 構築の経験をふまえて，NTCIR-2 について，プーリング法と追加検索による正解文書リストの作成が，相対的なシステム評価にどのような影響を与えるか考察する．

以下，2 章では，NTCIR ワークショップ 2 における，日本語・英語検索タスクのサブタスクと提出結果，および，プーリングによる正解文書リストの作成法について述べる．3 章では，プーリング実験と作成した数種類のプールを用いた評価実験について述べ，4 章では，本研究で分かったことをまとめる．

1.3 テストコレクションとプーリング法

大規模テストコレクションとしては，TREC（Text REtrieval Conference）[11)~13)] が有名である．TREC の登場以降，情報検索システムの評価実験は大規模テストコレクションを使用したものが主流となり，大規模なテストコレクションを使った実験を行わなければ，情報検索の国際的なコミュニティに受け入れられにくいというのが現状である．

テストコレクションを構築するとき，各検索課題について文書データベース中のすべての文書の正解判定を行い，完全に網羅的な正解文書リストを作成することが理想ではあるが，数万から数十万件の文書を含む大規模なデータベースに対してそのようなことを行うのは，実際には不可能であり，別の方法で正解文書を網羅的に収集する必要がある．

大規模テストコレクション構築における正解文書候

補の収集の手法としては，プーリング法[3]が，効率的で効果的な方法として有名であり，TREC では，1992年から，プーリング法によって大規模テストコレクションを作成している．

プーリング法では，異なる検索手法を用いた様々な検索システムによって検索された結果のそれぞれから，各検索課題ごとに上位一定数（X 件）ずつをプールし，プール中のすべての文書について人間の正解判定者が正解判定を行う．プールに含まれない文書は判定されず，不正解と仮定されるため，どのように，正解文書候補を網羅的に，また，どのような検索システムに対しても公平になるように集めるかが問題となる．

また，最近では，プーリング法の改良として Move-To-Front（MTF）プーリング法[2]が提案されている．MTF 法では，検索結果に優先順位を付け，優先順位の高い検索結果の文書を多くプールし，判定を行う．MTF プーリング法は，従来のプーリング法に比べて，より少ない正解判定で効率的に正解文書を作成することができるといわれている[2]．しかし，優先順位によって検索結果からプールする文書数を変えることは公平であるかどうかということは，検索システムの評価においては問題となる．

Zobel[14]は，TREC コレクションについて，(1) プーリングに参加しなかったシステムをプーリングで作成したテストコレクションで評価したときの影響，(2) 他のシステムと共通の多くの文書を検索してきたシステムの利点，(3) 平均精度と未判定文書数との相関，について実験と考察を行っている．その論文では，(1) ある検索システムにおいてのみ見つかった unique な正解文書を正解文書リストから削除した場合としなかった場合を比較するため，プールに参加したすべてのシステムについて unique な正解文書を除いたときの評価を行って平均をとったとき，11 点再現率–精度（11-point Recall-Precision）の平均は，プール数が 100 件の場合，unique な正解文書を除いたときの方が，TREC-5 では平均 0.5%，TREC-3 では平均 2.2%高くなり，プール数が 100 件より小さい場合には，より大きな差があった，(2) 全検索結果からのそれぞれプールする文書数を 50 件とし，ある 1 つの検索結果からだけさらに 50 件多くプールしたとき，11 点再現率–精度は，多くプールした検索結果では 1%向上し，他の検索結果では平均 0.5%下がった，(3) 検索結果の上位 100 件の unique な文書数とその検索結果中の未判定文書数とは相関があり，上位 100 件で unique な文書を多く見つけている検索結果（システム）は，それ以降の順位の未判定文書も unique であるが，未判定文書数と 11 点再現率–精度には相関がなかった，ということが報告されている．

このことから，検索結果からのプール数や検索結果の正解文書リストへの貢献度（uniqueness やプールに含まれたか含まれないか，など）が，プーリング法によって作成されたテストコレクションを用いた検索システムの評価に何らかの影響を与えるのではないかと考えられる．

本稿では，以上のようなことをふまえ，従来のプーリング法を用いて作成した大規模テストコレクション NTCIR-2 が，情報検索システムの評価ツールとして公平であるかどうか確かめるため，NTCIR ワークショップ 2 の提出結果を用いたプーリング実験と評価実験を行い，プーリング法によって作成した正解文書リストの網羅性と公平性が評価にどのような影響を与えるか考察する．

2. NTCIR-2 の正解文書リスト

大規模テストコレクション NTCIR-2 の正解文書リストは以下のような手順で作成された．(1) プーリング法を用いて正解文書候補を収集する，(2) 人間の正解判定者によって正解文書候補の正解判定を行う，(3) ある一定数以上の正解文書を持つ検索課題について，正解文書リストの網羅性を高めるため，対話型検索システムを用いて，再現率を重視した検索を行い，追加の正解文書候補を収集する，(4) 追加の正解文書候補について正解判定を行う．次節以下で，サブタスクと正解文書作成手順の各ステップについて詳しく説明する．

2.1 日本語・英語検索タスクのサブタスクと検索対象文書

2.1.1 サブタスク

本項以下では，NTCIR ワークショップ 2[10]の「日本語・英語検索タスク（Japanese & English IR Task）」を「JEIR タスク」と略記する．JEIR タスクには 2 つのサブカテゴリがあり，そのサブカテゴリ中のサブタスクは以下のとおりである．

単言語検索タスク：

- J-J タスク：日本語検索課題を用いて日本語文書を検索する，
- E-E タスク：英語検索課題を用いて英語文書を検索する．

言語横断検索タスク：

- E-J タスク：英語検索課題を用いて日本語文書を検索する，
- J-E タスク：日本語検索課題を用いて英語文書を検索する，

- J-J,E タスク：日本語検索課題を用いて日本語文書と英語文書を検索する，
- E-J,E タスク：英語検索課題を用いて日本語文書と英語文書を検索する．

2.1.2 サブタスクにおける検索対象文書

JEIR タスクでは，J コレクションと E コレクションという 2 つの文書コレクションが使用された．J コレクションと E コレクションは，国立情報学研究所（NII）が提供している「学会発表データベース」と「科学研究費補助金研究成果概要データベース」の一部を抽出したものである．元のデータベースの一部は，日本語文書と英語文書が対訳として組になっている．

J コレクションは，*ntc1-j1.mod*，*ntc2-j1g*，*ntc2-j1k* という 3 つの文書セットから成り，E コレクションは，*ntc1-e1.mod*，*ntc2-e1g*，*ntc2-e1k* という 3 つの文書セットから成る．*ntc1-j1.mod*，*ntc1-e1.mod*，*ntc2-j1g*，*ntc2-e1g* は「学会発表データベース」から抽出された文書セット，*ntc2-j1k* と *ntc2-e1k* は「科学研究費補助金研究成果概要データベース」のから抽出された文書セットである．

日本語文書と英語文書が元のデータベース中で 1 組の対訳になっているとき，それらは同じ文書番号（ACCN）が付与されている．NTCIR-2 では，J コレクションと E コレクションを独立な文書コレクションとして扱うため，組になっている日本語文書と英語文書を分け，英語文書には新たな ACCN を付与した．たとえば，元の文書の ACCN が「gakkai-000040700」であるとき，英語文書には，新たな ACCN「gakkai-e-000104007」を付与した．

各文書コレクションの文書数と，文書コレクション中に含まれる，元のデータベースでは組になっている文書数を**表 1** に示す．また，JEIR タスクのサブタスクと，NTCIR-2 の検索課題，検索対象文書，および正解文書リストのファイル名との対応を**表 2** に示す．

表 1 日本語・英語検索タスクで使用された文書コレクションと文書数
Table 1 Number of documents in the Document Collections used for the 2nd NTCIR Workshop.

文書コレクション	文書数
ntc1-j1.mod	332,918
ntc1-e1.mod	187,080
pairs in ntc1-j1&e1	181,485
ntc2-j1g	116,177
ntc2-e1g	77,433
pairs in ntc2-j1g&e1g	74,180
ntc2-j1k	287,063
ntc2-e1k	57,545
pairs in ntc2-j1k&e1k	57,510

2.1.3 提出結果からのプーリング

JEIR タスクのサブタスクの参加チームは，各自の検索システムを用いて，各検索課題について，J コレクションあるいは（および）E コレクションを検索し，検索結果を提出する．以下では，提出された検索結果を「run」と呼ぶ．

図 1 に JEIR タスクのプーリングの過程を示す．

JEIR タスクのプーリングでは，(1) すべてのサブタスクの run について，上位 X 件の文書をプールし，(2) プール中の日本語文書と英語文書の ACCN を，元の ACCN に戻す．たとえば，日本語文書の ACCN「gakkai-j-000040700」と英語文書の ACCN「gakkai-e-000104007」をそれぞれ元の ACCN「gakkai-000040700」（数字部分は日本語文書の文書番号と同一）に戻し，英語文書を日本語文書に対応付ける．この過程は，タスク横断的・言語横断的であるので，本稿では，言語横断的プーリングという言葉で表す．

プーリングは，すべての run について行ったわけではなく，2 つの理由から，いくつかの run はプーリングには使用しなかった（その理由については，付録 A.1 を参照）．**表 3** に実際のプーリングに使用した run の個数を示す．

各 run からプールされる文書数 X は 1 つの検索課

表 2 サブタスク，検索課題，文書，正解文書リストの関係
Table 2 Relationship of Sub-tasks, Search Topics, Documents, and Relevance Judgments.

タスク	検索課題	文書コレクション	正解文書リスト	
			Level1 (S or A)	Level2 (S, A or B)
J-J	topic-j101-149	ntc1-j1.mod, ntc2-j1g, ntc2-j1k	rel1_ntc2-j2_0101-0149	rel2_ntc2-j2_0101-0149
E-J	topic-e101-149			
J-E	topic-j101-149	ntc1-e1.mod, ntc2-e1g, ntc2-e1k	rel1_ntc2-e2_0101-0149	rel2_ntc2-e2_0101-0149
E-E	topic-e101-149			
J-J,E	topic-j101-149	ntc1-j1.mod, ntc2-j1g, ntc2-j1k, ntc1-e1.mod, ntc2-e1g, ntc2-e1k	rel1_ntc2-je2_0101-0149	rel2_ntc2-je2_0101-0149
E-J,E	topic-e101-149			

topic-j101-149 は日本語検索課題のリスト，topic-e101-149 は英語検索課題のリスト．

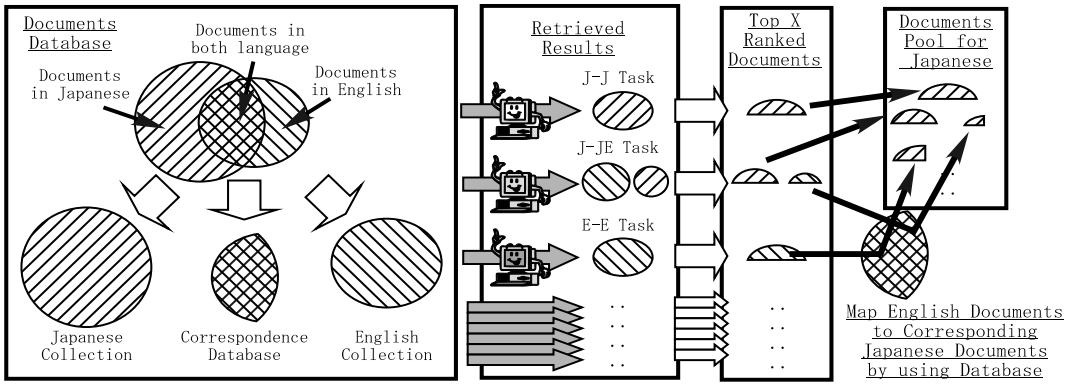

図 1 言語横断的プーリングの過程
Fig. 1 Process of Cross-Lingual Pooling.

表 3 提出された run の数とプールに使用された run の数
Table 3 Number of submitted runs and pooled runs.

タスク	提出された run	プールされた run
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

J-E タスクと J-J,E タスクの括弧内の数字 n は，JEIR タスクのオーガナイザの提出した run の数である．

題に対してプールされた総文書数が 2,000 から 2,500 件程度になるように検索課題ごとに 70，80，90，100 のいずれかに調整した．プールされた正解文書候補の正解判定は，人間の正解判定者が行うが，経験的に，1 つの検索課題について文書間で正解になるべく矛盾がないように正解判定を行える上限は 2,000 件程度であると考えられるため，そのような調整を行った．同一の検索課題については，どの run からも同じ上位 X 件をプールしている．

プーリングの過程で，ステップ (1) で作成されたプールのうち，日本語文書部分を $J1$，英語文書部分を $E1$，ステップ (2) で $J1$ から変換された英語文書を $E1$ に追加したプールを $E2$，$E1$ から変換された日本語文書を $J1$ に追加したプールを $J2$ とすると，それぞれに含まれる正解文書数の，最終正解文書リスト F のうちの単言語の正解文書リスト J(F)，E(F) に対する割合は，$J1$:89.6%，$E1$:91.8%，$J2$:96.6%，$E2$:98.1%となっている．このことから，言語横断的プーリングが，対訳になっている文書の中で新たな正解文書を見つけるのにかなり効果的であることが分かった．

2.2 正解判定

正解判定作業では，検索課題ごとに，主判定者 1 人と副判定者 1 人の計 2 人の正解判定者が正解判定とクロスチェックを行った．主判定者は，基本的には，その検索課題の作成者である．最終判定結果は，2 人の判定者の協議に基づき，主判定者が決定した．正解判定は，高正解（高適合）highly-relevant (S)，正解 relevant (A)，部分的正解 partially-relevant (B)，不正解 non-relevant (C) の 4 つのレベルで行った．NTCIR ワークショップ 2 では,「S」と「A」を正解,「B」と「C」を不正解とした正解文書リスト（Level1）と,「S」,「A」,「B」を正解とし,「C」を不正解とした正解文書リスト（Level2）を用いて評価を行っているが，本稿では，Level2 の正解「S」,「A」,「B」を「正解文書」として定義し，以下では，その意味で使用する．

検索課題を作成する時点で，検索課題作成者には，対話型検索システムを用いて予備的な検索を行ってもらい，5〜10 件の正解文書候補のリストを提出してもらった．この予備検索結果文書のうち，参加者の run から作成したプールを P とし，P に含まれていない，予備検索だけで見つかった文書の集合をプール PP とする．正解判定は，P と PP を合わせた正解文書候補リスト $P+PP$ に対して行い，全 49 件の検索課題のうちの 29 件の検索課題についてはクロスチェックを行った．最後に，すべての検索課題について，各主判定者が最終的なチェックを行い，最終判定結果を決定した．

2.3 追加検索

参加者からの run のプールについての正解判定後，正解文書を 110 件以上持つ検索課題あるいは各 run について上位 70 件の文書をプールした検索課題 16 件については，正解文書リストの網羅性を高めるため，

表 4 プール中の正解文書数の割合
Table 4 Number of relevant documents in the pools.

平均	J(P)	J(PP)	J(I)	J(F)	E(P)	E(PP)	E(I)	E(F)
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

ave%16 追加検索を行った 16 件の検索課題についての F に対する割合の平均であり，ave%all は，全検索課題についての平均である．

対話型検索システムを用いて，図書館情報学専攻の大学院生によって，再現率を重視した追加検索を行った．そして，その追加検索で新たに見つかった文書の集合 I について主判定者によって追加判定を行い，I を $P+PP$ に加えて最終的な正解文書リスト F を作成した．

run からのプール P，予備検索のプール PP，追加検索のプール I，の正解文書数の，最終正解文書リスト F のうちの日本語部分 J(F) と英語部分 E(F) に対する検索課題ごとの割合の平均を**表 4** に示す．ave%16 は，追加検索を行った 16 件の検索課題についての平均，ave%all は，全検索課題についての平均である．

表 4 から分かるように，正解文書を 110 件以上持つ，あるいは，プール数が 70 件であった検索課題 16 件についての J(P) と E(P) の，それぞれ J(F) と E(F) に対する割合の平均は，91.4%と 95.3%であり，NTCIR ワークショップ 1 の予備テストと評価テストでの run からだけのプールに含まれる正解文書の割合よりもかなり大きい．これは，NTCIR ワークショップ 2 では，NTCIR ワークショップ 1 よりも多くの run が提出されたため，run からのプールの網羅性が高まったからだと考えられる．しかしながら，再現率重視の追加の対話型検索は，日本語の正解文書 J(F) の 8.4%を見つけており，ある程度，網羅性を高めるのに効果的であったといえる．

3. プーリングおよび評価の実験

3.1 実験の目的

プーリングによって作成された正解文書リストの網羅性と公平性がシステム評価にどのような影響を与えるのか調べるため，以下の 3 つの点について，プーリング実験とプールに入れた run の評価を行った．

3.1.1 追加検索の影響

Zobel [14] と Voorhees [12] らは，正解文書が多い検索課題については，網羅性の点から，プールの深さ（pool depth），すなわち，各 run から上位何件を一定数としてプールするかが問題である，と言っているが，一方では，正解である可能性の高い文書を検索結果から個別に追加する Cormack ら [2] の MTF プーリング法については，多くの文書をプールされた検索結果の方が評価の際に有利になる可能性があり，評価が公平ではなくなるのではないか，ということを指摘している．筆者らは，このことを考慮し，NTCIR-1 と 2 の構築においては，正解文書リストの網羅性を高めるために，検索結果から追加で文書をプールするかわりに，正解文書を多く持つ検索課題については再現率を重視した人間の検索者による対話型の追加検索を行い，正解文書を補完した．これは，他のテストコレクション構築では行われていない試みである．この追加検索が検索システムの評価に与える影響ついて，筆者らの論文 [4)~6)] で述べたように，NTCIR-1 に関しては，平均精度の平均による相対的評価には影響がないということが確かめられた．本稿では，NTCIR-2 によるシステム評価に関しても，同様に実験を行って考察する．

3.1.2 プーリングに参加しなかったシステムの評価への影響

1.2 節で述べたように，Zobel [14] は，TREC-3 と 5 について，正解文書リストから各検索システムの unique な文書を削除して評価実験を行うことによって，プーリングに参加しなかったシステムの評価をシミュレーションしている．本稿でも，この論文に倣い，プーリングに参加しなかったシステムへの影響を検証する．

3.1.3 言語横断プーリングのサブタスクごとの影響

NTCIR-2 では，NTCIR-1 では行わなかった新しい試みとして，言語横断的プーリングによる正解文書リストの作成を行っている．2.1.3 項で述べたように，その過程で，プーリングの効率を下げると考えられる J-J,E タスク，E-J,E タスクの run は（その run しか全タスクを通して提出されていない場合を除いて）プールには使用しなかった．そのことが正解文書リストにどのような影響を与えるのか調べるため，サブタスクごとのプールを作成し，サブタスクごとの正解文書リスト，各サブタスクでの unique な正解文書を除いた正解文書リストをそれぞれ作成し，評価実験を行う．

3.2 プーリング実験

3.2.1 追加検索について

対話型検索システムを用いた追加検索がシステム評価に影響を与えるかどうか調べるために，J-J タスクの run について，最終的な正解文書リスト F と，F から対話型検索でのみ見つかった文書集合 I を除い

表 5 サブタスクごとのプール中の正解文書数の割合
Table 5 Number of relevant documents in the pools for sub-tasks.

平均	J($P(J-J)$)	J($P(J-E)$)	J($P(E-E)$)	J($P(E-J)$)
ave % F	81.6	35.3	31.9	73.5
平均	E($P(J-J)$)	E($P(J-E)$)	E($P(E-E)$)	E($P(E-J)$)
ave % F	72.9	86.5	78.8	67.3
平均	J($F-P(J-J)$)	J($F-P(J-E)$)	J($F-P(E-E)$)	J($F-P(E-J)$)
ave % F	87.8	97.3	98.6	94.4
平均	E($F-P(J-J)$)	E($F-P(J-E)$)	E($F-P(E-E)$)	E($F-P(E-J)$)
ave % F	97.4	93.3	96.9	98.5

た文書リスト $F-I$ を正解文書リストとして用いて評価実験を行った．

3.2.2 各システムの unique contribution について

全タスクについて，同じ検索システムから提出された run だけが見つけた正解文書をそのシステムの unique contribution S とし，最終的な正解文書リスト F から S を除いた文書リスト $F-S$ をシステムごとにそれぞれ作成し，J-J タスクの各 run について，$F-S$ を正解文書リストとして用いた評価実験を行った．

3.2.3 サブタスクごとのプールについて

サブタスクごとの run からのプールが，正解文書リストの網羅性にどれくらい貢献しているか，また，相対的なシステム評価に影響を与えるいるのかどうか調べるため，サブタスク J-J，E-J，E-E，E-J について，サブタスクごとのプール $P(task)$ と，最終正解文書リスト F から，各サブタスクのプールだけに含まれる文書の集合を除いたプール $F-P(task)$ を作成し，評価実験を行った．

その各プール中の正解文書数の正解文書全体に対する割合の平均を**表 5** に示す．

3.3 評価実験

最終的な正解文書リスト F，F から対話型検索でのみ見つかった文書集合 I を除いた文書リスト $F-I$，各システムの unique contribution を除いたプール $F-S$，サブタスクごとのプール $P(J-J)$，$P(J-E)$，$P(E-E)$，$P(J-E)$，F からサブタスクごとのプール中のユニークな文書を除いたプール $F-P(J-J)$，$F-P(J-E)$，$F-P(E-E)$，$F-P(E-J)$ を，それぞれ正解文書リストとして用いて評価を行った．追加検索についてのプール，サブタスクについてのプールを正解文書リストとして用いて評価を行ったときの，J-J タスクの run の平均精度の平均とそれによる順位を**表 6** に示す．また，各システムの unique contribution を除いたプールを正解文書リストとして評価したときの，そのシステムの run の平均精度の平均と，最終的な正解文書 F による平均精度の平均との差を**表 7** に示し，平均精度の平均のグラフを**図 2** に示す．

表 6 には，評価に使用した run の Run-ID，その run が検索に使用した検索課題のフィールド，検索式作成方法，各正解文書リストによるその run の平均精度の平均に基づく順位とその平均精度の平均を示す．表 7 には，評価に使用した run の Run-ID，その run が検索に使用した検索課題のフィールド，検索式作成方法，最終的な正解文書 F によるその run の平均精度の平均，各システムの unique contribution を除いたプールを正解文書リストとして評価したときの，そのシステムの run すべての平均精度の平均，および F による平均精度の平均との差を示す．

テストコレクション NTCIR-2 の検索課題は，5 つのフィールド，「TITLE（T）」，「DESCRIPTION（D）」，「NARRATIVE（N）」，「CONCEPT（C）」，「FIELD（F）」から成っている（検索課題の詳しい形式については，付録 A.2 を参照）．評価には，提出された全 run のうち，検索に使用した検索課題のフィールドが「DESCRIPTION（D）」である run を参加チームごとに 1 つずつ用いた．DESCRIPTION を使用した run を提出していないチームについては，使用フィールドに DESCRIPTION を含む run をかわりとして用いた．表 6 と表 7 では，検索フィールドの頭文字 1 文字を用いて，その run の使用フィールドを示す．たとえば，「D」はその run が「DESCRIPTION」のみを使用したということ，「D N C」はその run が「DESCRIPTION」と「NARRATIVE」と「CONCEPT」を使用したということである．

検索式作成法は，検索式の作成に人間が関わるか関わらないかによって，対話型（手動）（interactive）と自動（automatic）に大きく分けられる．表 6 と表 7 では，それぞれ，inter，auto と略記する．

表 6 では，Run-ID は，最終正解文書リスト F によって評価したときの，その run の順位の順に左から並んでいる．一般的に，類似した検索システムによる異なる run の平均精度の平均を比較する場合には，その差が 5〜7%であるとき，統計的に有意であり，同じシステムによる異なる run を比較する場合には，そ

表 6 J-J タスクの run の平均精度の平均と順位
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
使用フィールド	D N	D	D	D	D	D	T D N C F	D	D	D
検索式作成法	interact	auto	auto	auto	auto	auto	interact	interact	auto	auto
F	1	2	2	4	5	5	7	7	9	9
	0.4138	0.3686	0.3620	0.3377	0.3051	0.3024	0.2847	0.2810	0.2687	0.2683
F-I	1	2	2	4	5	5	7	7	9	9
	0.4173	0.3720	0.3659	0.3396	0.3085	0.3059	0.2863	0.2814	0.2713	0.2713
P	1	2	2	4	5	5	7	7	9	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	2	4	5	5	7	7	9	9
	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	3	5	**7**	**6**	**11**	**8**	**8**
	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	3	5	**7**	**6**	**11**	**8**	**8**
	0.1951	0.1801	0.1682	0.1652	0.1512	0.1354	0.1464	0.1177	0.1261	0.1271
P(E-J)	1	2	2	4	5	5	7	7	7	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	2	4	5	5	7	7	7	7
	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	2	4	5	5	7	7	7	7
	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	2	4	5	5	7	8	9	9
	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	2	4	5	5	7	8	9	9
	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
使用フィールド	D	D	D	D	D	D N C	D	D	D	D
検索式作成法	auto	auto	auto	auto	auto	interact	auto	auto	auto	auto
F	11	12	13	14	14	14	14	18	19	20
	0.2567	0.2432	0.2309	0.2101	0.2067	0.2059	0.2016	0.1591	0.1210	0.0138
F-I	11	12	13	14	14	14	14	18	19	20
	0.2579	0.2454	0.2329	0.2121	0.2093	0.2076	0.2044	0.1599	0.1205	0.0141
P	11	12	13	14	14	14	14	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	14	14	14	14	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)	11	**8**	13	13	**16**	**17**	**13**	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**	13	13	13	**17**	**12**	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	11	13	14	14	16	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	11	13	14	14	14	14	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	12	14	14	14	14	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	12	14	14	14	14	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	11	13	14	14	14	14	18	19	20
	0.2642	0.2513	0.2389	0.2173	0.2156	0.2120	0.2086	0.1637	0.1235	0.0145

使用フィールドは，その run の検索に使用された検索課題中のフィールドである．

の差が 1〜7%であるとき，有意であるといわれている[1),14)]．本稿では，隣り合う run の平均精度の平均の差が 5%以下であるときには，同順位として表す．各正解文書リストによる各 run の順位が入れ替わっている場合には，その順位の数値を太字にし下線を引く．

表 6 の第 4 行（F による順位）と第 6 行（$F-I$ による順位）から，F，$F-I$ を用いて評価した結果，各 run の順位はまったく同じになることが分かる．特に，最終正解文書リスト F で評価しても，F から追加の対話型検索だけが見つけた文書集合 I を除いたリスト $F-I$ で評価しても，対話型手法を用いている 4 つの run，DOVE9，FXSD2，sstut6，smlab のいずれの順位も変わっていない．このことから，追加検索は，正解文書リストの網羅性の向上には貢献しているが，システム評価に与える影響は無視できる程度であることが分かった．

また，表 6 の第 4 行（F による順位）と第 8，10，12，14，17 行（各サブタスクことのプールによる順位）から，F とその他のサブタスクごとのプールを用いた評価による順位も，ほとんど変わらないことが分かる．英語文書を検索対象としたサブタスク J-E タスクと E-E タスクからのプールについては，第 14 行

表 7 各システムの unique contribution を除いたときの J-J タスクの run の平均精度の平均と差分
Table 7 Mean average precision and difference of each run for the J-J task without each run's unique contribution.

Run-ID	DOVE9	CRL16	LAPIN6	JSCB1	R2D22	sstut1	FXSD2	sstut6	aplij2	DOVE3
使用フィールド	D N	D	D	D	D	D	T D N C F	D	D	D
検索式作成法	interact	auto	auto	auto	auto	auto	interact	interact	auto	auto
F	0.4138	0.3686	0.3620	0.3377	0.3051	0.3024	0.2847	0.2810	0.2687	0.2683
F-S	0.4173	0.3735	0.3653	0.3414	0.3084	0.3038	0.2845	0.2831	0.2740	0.2719
	0.9(%)	1.3(%)	0.9(%)	0.4(%)	1.1(%)	0.5(%)	-0.1(%)	0.7(%)	2.0(%)	1.3(%)

Run-ID	FXSD1	Brkly2	SRGDU1m	STIX6	MP1NS5	smlab	sato2	WUSKL	OASIS9	trans4
使用フィールド	D	D	D	D	D	D N C	D	D	D	D
検索式作成法	auto	auto	auto	auto	auto	interact	auto	auto	auto	auto
F	0.2567	0.2432	0.2309	0.2101	0.2067	0.2059	0.2016	0.1591	0.1210	0.0138
F-S	0.2590	0.2480	0.2339	0.2123	0.2095	0.2039	0.2050	0.1605	0.1224	0.0139
	0.9(%)	2.0(%)	1.3(%)	1.0(%)	1.4(%)	-1.0(%)	1.7(%)	0.9(%)	1.2(%)	0.7(%)

使用フィールドは，その run の検索に使用された検索課題中のフィールドである．
本表では，表 6 と異なり，S は列ごとに異なっている．S はその列の run を提出したシステムであり，すなわち，たとえば，run が DOVE9 であるときには，S =DOVE であり，run が CRL16 であるとき，S =CRL である．

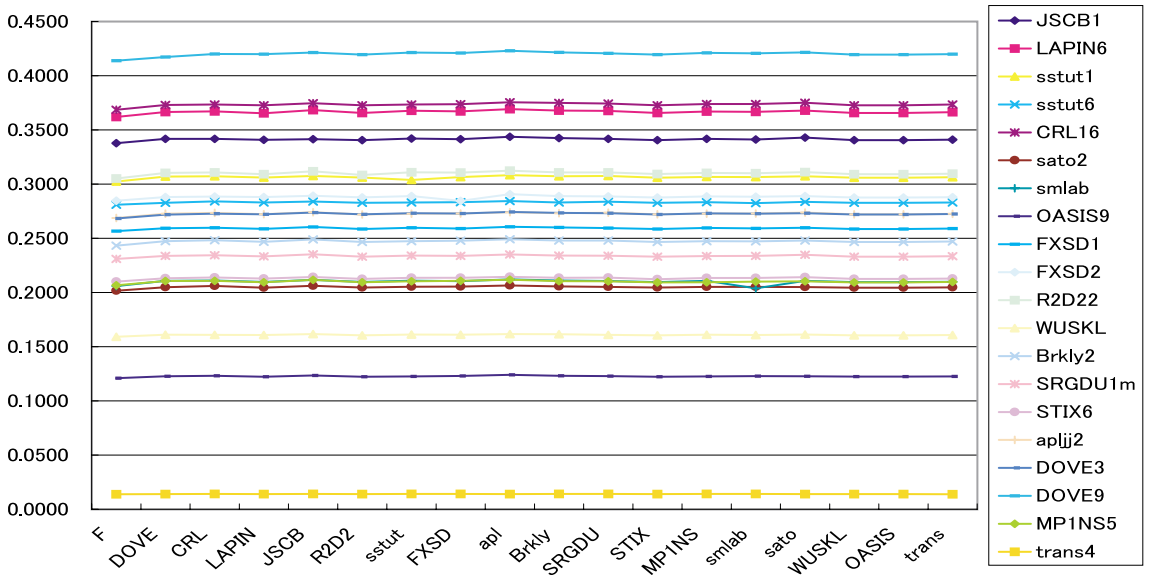

図 2 各システムの unique contibution を除いたときの J-J タスクの run の平均精度の平均
Fig. 2 Mean average precision of each run for the J-J task without each run's unique contribution.

($P(J-E)$ による順位) と第 17 行 ($P(E-E)$ による順位) から，中程度の順位の run の順位は入れかわっているが，この 2 つのプールによる評価結果の傾向は似ていることが分かる．

この傾向は，J-E タスクと E-E タスクでは，run は E コレクション中の英語文書を検索対象としているため，英語文書の文書番号 ACCN を日本語文書の ACCN に対応付けすることによって，正解文書候補として対応する日本語文書も正解文書候補としてプールすることができるが，英語と対応している日本語文書は，J コレクション全体の中では部分集合であるので，J-J タスクと E-J タスクで使用される日本語文書全体に対する正解文書リストの中の一部にすぎないため，J-E タスクあるは E-E タスクのみからのプールでは，正解文書の分布が偏っている可能性あることに由来しているのではないかと考えられる．さらに，J-E タスクと E-E タスクのぞれぞれの run からだけのプールの網羅性は，J-J タスクと E-J タスクのそれぞれからのプールよりも低いので，J-J タスクの run を，J-E タスクの run からだけのプール $P(J-E)$ を正解文書リストとして使用して評価すると，J-J タスクからの run のプール $P(J-J)$ を正解文書リストとして評価したときよりも，平均精度の平均が低くなる．それは，プール $P(J-E)$ に含まれていない文書は，評価の際，不

正解と仮定されるため，評価される各 run に含まれる不正解の件数が増えるからである．$P(E-E)$ を正解文書リストとして使用した場合にも同様である．

したがって，異なる言語の文書コレクションの文書の対応関係がアンバランスであるとき，文書数が少ない方の 1 言語の文書コレクションを検索対象とするサブタスクからのみのプーリングでは，正解文書を網羅的に収集できない可能性があることが分かる．1 つの言語の文書コレクションのみからのプーリングは正解文書の網羅性に影響を与えるが，複数の対訳文書コレクションについての言語横断的プーリングは，正解文書収集の網羅性を高める効果があり，また，システム評価にほとんど影響を与えないと考えられる．

表 7 の第 6 行（F による評価と $F-S$ による評価の差分）と図 2 を見ると，あるシステムだけが見つけた正解文書を正解文書リストから除いた場合，そのシステムの提出した run の平均精度の平均は，平均で 1.1%，最大で 2.4%高くなり，そのシステムの run の評価にはある程度影響があるが，そのシステムの unique contribution があってもなくても，平均精度の傾向はあまり変わらないことがが分かる（平均精度の平均が高くなるのは，unique な正解文書を除くと正解文書数全体が減るためである）．したがって，そのシステムの見つけた unique な正解文書を除くことは，評価値の絶対値には影響を与えるものの，平均精度の差は 5%以下に収まっているので，unique contribution の有無は相対的な評価には，ほとんど影響がないと考えられる．このことにより，あるシステムが，プーリングに参加していなかったとしても，そのシステムの評価にプーリングによって作成したテストコレクションを使用することは可能であることが分かった．

4. おわりに

言語横断的プーリングと，対話型検索システムによる再現率重視の追加検索を用いて作成された正解文書リストの公平性を調べるために，NTCIR ワークショップ 2 の日本語・英語検索タスクの評価テストの提出結果（run）を用いて，実験的なプーリングと評価を行った．実験によって，次のようなことが分かった．

(1) 追加の対話型検索のシステム評価への影響：追加検索は，正解文書を 110 件以上持つ検索課題および上位 70 件だけをプールした検索課題との合計 16 件について行われた．最終的な正解文書リスト F と，F から追加検索で見つかった正解文書 I を除いた $F-I$ を用いて，各 run を評価し，検索課題全体に対する平均精度の平均で順位付けを行った結果，各 run の相対的な順位はまったく変わらなかった．この実験的な結果は，対話型検索システムによる追加検索を行っても，システム評価には影響がない[4)~6)]という仮定を補強するものである．

(2) 追加検索の必要性：追加検索は，正解文書リストの網羅性を高めるのにある程度有効であるが，プーリングに使用できる run の個数が十分多く，多様であれば，追加検索は必要ではないかもしれない．しかし，何件の run をプーリングに使用すれば十分であるかは不明であるので，その件数と多様性について検討する必要がある．

(3) プーリングに参加しなかった検索システムの評価への影響：プーリングに参加しなかった検索システムの評価をシミュレートするため，各システムだけが見つけた正解文書を正解文書リストから除き，そのシステムの run の評価を行った．最終的な正解文書リストを用いた評価との差は，平均で 1.1%，最大で 2.4%であったが，相対的な評価として考えるときには，ほとんど差はなく，システム評価にはほぼ影響がないことが分かった．したがって，プーリングに参加しなかったシステムを評価するツールとしても，NTCIR-2 は有効であると考えられる．

(4) 言語横断的プーリングのシステム評価への影響：サブタスクごとのプールを作成し，そのプールを用いて各 run を評価したとき，各プールがどれくらい正解文書リストの網羅性に貢献しているかは異なるので（表 5 参照），各 run の平均精度の平均値の絶対的な大きさは異なる（表 6 参照）．しかし，どのプールを用いても，検索課題全体に対する平均精度の平均での相対的な順位は，最終正解文書リスト F による順位とほとんど変わらなかった．結果として，言語横断的プーリング，すなわち，単言語文書の検索結果からプールした正解文書候補を，異なる言語の文書における正解文書候補として加えたことは，システム評価にほとんど影響を与えず，また，正解文書候補を効率的に集めるためにある程度有用であることが分かった．

謝辞 システム評価について貴重なご助言をしてくださった，駿河台大学岸田和明助教授に深く感謝いたします．

本研究は，日本学術振興会未来開拓学術研究推進事業「高度分散情報資源活用のためのユビキタス情報システム」（課題番号 JSPS-RFTF96P00602）による．

参 考 文 献

1) Buckley, C. and 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, *Proc. ACM–SIGIR'98*, Melbourne, pp.282–289 1998.
3) Gilbert, G. and Sparck Jones, K.: Statistical Bases of Relevance Assessment for the 'Ideal' Information Retrieval Test Collection, BL R&D Report 5481 (1979).
4) 神門典子ほか：NTCIR-1：情報検索システム評価用テストコレクション構築の方針と実際，情報処理学会研究報告，99–FI–53–5, pp.33–40 (1999).
5) 栗山和子ほか：大規模テストコレクション構築のためのプーリングについて：NTCIR-1 の予備テストの分析，情報処理学会研究報告，99–FI–54–4, pp.25–32 (1999).
6) 栗山和子ほか：大規模テストコレクション NTCIR-1 の構築（1）：プーリングと正解判定の分析，情報処理学会第 59 回全国大会，pp.3-105–3-106 (1999).
7) 栗山和子ほか：大規模テストコレクション構築のためのプーリングについて：NTCIR-1 の分析，学術情報センター紀要，pp.17–38 (2000).
8) NII-NACSIS Test Collection for Information Retrieval systems. http://research.nii.ac.jp/ntcir/
9) 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.
10) NTCIR Workshop 2: *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, Mar.7–9 (2001), ISBN 4-924600-89-X.
11) Text REtrieval Conference (TREC). http://trec.nist.gov/ (visited January 12th, 2001).
12) Voorhees, E.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, *Proc. ACM–SIGIR'98*, Melbourne, pp.315–332 (1998).
13) Voorhees, E. and Harman, D. (Eds.): The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-242, Maryland, U.S.A. (2000).
14) Zobel, J.: How Reliable are the Results of Large Scale Information Retrieval Experiments?, *Proc. ACM–SIGIR'98*, Melbourne, pp.307–314 (1998).

付　　録

A.1 いくつかの run をプーリングに使用しなかった理由

プーリングは，すべての run について行ったわけではなく，次のような理由から，いくつのか run はプーリングには使用しなかった．提出された run の個数と実際のプーリングに使用した run の個数は，表 3 に示したとおりである．

A.1.1 プールする文書数上位 X 件の公平性

表 2 から分かるように，J-J,E タスクと E-J,E タスクでは，J コレクションと E コレクションを使用している．E コレクションの文書は，J コレクションの文書の一部と対訳になっているため，ACCN を変換した後，重複した ACCN の文書を除くと，J-J,E タスクと E-J,E タスクの 1 つずつの run からプールされる文書数は実際は上位 X 件よりも少なくなり，プーリングの効率が落ちる．J-J,E タスクと E-J,E タスクの run についてだけ，重複を除いてから上位 X までプールするとすれば，効率的なプーリングは行えるが，実際には，1 つの run から $X + \alpha$ をプールすることになり，α は同じサブタスクの中でも run によって異なり，各 run から上位一定数をプールするという公平なプールの原則が崩れてしまう．

効率的なプーリングを行うため，J-J,E タスクと E-J,E タスクの run については，同じ検索システムを用いた同じ参加チームの run が，他の 4 つのタスクのうちのいずれにも含まれていない場合を除いて，プーリングに使用しないことにした．すなわち，タスク全体の run の中で J-J,E タスクあるいは E-J,E タスクのみでしか結果を提出していない参加チームの run 以外は，J-J,E タスクと E-J,E タスクの run はプーリングに使用しなかった．

A.1.2 同一システムの提出結果のプールの制限

1 つの参加チームが同じ検索システムを用いた run を複数提出している場合，パラメータなどの違いによって各 run は異なっているけれども，まったく異なるシステムの run よりは，比較的同じような文書を見つけてくるのでないか，という仮定から，効率的なプーリングと正解判定のために，参加チームが提出時に run に付けた優先順位に基づいて，1 つの参加チームにつき，タスク別に，優先順位の高い順に 2 つまでの run をプーリングに使用した．

```
〈TOPIC q=0101〉

〈TITLE〉
B 型肝炎
〈/TITLE〉

〈DESCRIPTION〉
遺伝子工学的手法による B 型肝炎ワクチンの開発について論じている文献
〈/DESCRIPTION〉

〈NARRATIVE〉
肝炎などのウイルス性疾患に対する安全かつ有効な予防法の確立は 21 世紀に向けての医療分野での重要な課題である．そのため，遺伝子工学的手法による B 型肝炎ワクチンの開発について論じていれば検索要求を満たす．開発された B 型肝炎ワクチンの物理化学的特性を論じているものやその免疫力増強に有効な免疫アジュバントについて論じているものも検索要求を満たす．しかし，遺伝子工学的手法に触れていない論文は不可．また，B 型肝炎以外のワクチンも不可．
〈/NARRATIVE〉

〈CONCEPT〉
a. B 型肝炎,
b. 遺伝子工学的手法,
c. ワクチン，予防接種
〈/CONCEPT〉

〈FIELD〉
7. 医学・歯学
〈/FIELD〉

〈/TOPIC〉
```

図 3 NTCIR-2 の検索課題の例
Fig. 3 A sample of search topics in NTCIR-2.

A.2 検索課題の形式

A.2.1 検 索 課 題

検索課題は，分野の専門家（大学院生以上）から，インタビューあるいは一定の形式で形式の検索課題記入フォームによって収集された[4]．正解文書数が少なすぎて検索性能評価に影響を与えないように，予備検索を行い，正解文書数が 5 件以上あるものを選択している．

検索課題の形式は，初期の TREC の検索課題に準じ，SGML 形式に類似したタグが付与されている（タグの詳細は，文献 4) を参照）．

NTCIR ワークショップ 2 では，検索システムの訓練用としては，NTCIR-1 の検索課題 83 件を使用し，評価用検索課題として新たに日本語検索課題と英語検索課題それぞれ 49 件を用意した．日英の検索課題は対訳になっている．

図 3 に NTCIR ワークショップ 2 の評価用検索課題の例を示す．

(平成 13 年 9 月 25 日受付)
(平成 14 年 1 月 7 日採録)

(担当編集委員 岩山 真)

栗山 和子（正会員）

1993 年図書館情報大学大学院図書館情報学研究科修了．1996 年筑波大学大学院工学研究科博士課程修了．博士（工学）．同年，同大学準研究員．1998 年学術情報センター（現，国立情報学研究所）リサーチ・アソシエイト．2001 年国立情報学研究所 COE 研究員，現在に至る．数式処理，情報検索の研究に従事．日本数式処理学会，日本応用数理学会，ACM（SIGSAM，SIGIR）各会員．

吉岡 真治（正会員）

1991 年東京大学工学部精密機械工学科卒業．1996 年同大学大学院博士課程修了．博士（工学）．同年学術情報センター助手．2000 年国立情報学研究所助手．2001 年北海道大学大学院工学研究科助教授，現在に至る．知識ベースに基づく設計支援システムの構築，設計過程のモデル化，専門用語解析のためのコーパス作成等の研究に従事．人工知能学会，日本機械学会，精密工学会，ASME 各会員．

神門 典子（正会員）

1994 年慶應義塾大学文学研究科博士課程修了．博士（図書館・情報学）．同年学術情報センター助手．1995 年米国シラキウス大学情報学部客員研究員，1996〜1997 年デンマーク王立図書館情報大学客員研究員．1998 年学術情報センター助教授．2000 年国立情報学研究所助教授，現在に至る．テキスト構造を用いた検索と情報活用支援，言語横断検索，情報検索システムの評価等の研究に従事．ACM-SIGIR，BCS-IRSG，ASIS&T，言語処理学会，日本図書館情報学会各会員．