2003－FI－ 72　(13)
2003－NL－157　(13)
2003／9／30

大規模テストコレクション構築のためのプーリング：NTCIR-3 言語横断検索タスクの分析

栗山和子* 江口浩二† 岸田和明‡ 神門典子§

概要. 大規模テストコレクション NTCIR-3 の言語横断検索システム評価用データの適合文書リストは、NTCIR ワークショップ 3 の言語横断検索タスク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。本研究では、NTCIR-3 の作成過程において行なわれた、サブタスク混合プーリングが適合文書リストの網羅性を高めるのに有効であったかどうかについて考察する。
NTCIR ワークショップ 3 の提出結果を用いて、日本語文書についてのプーリング実験を行なった結果、単言語検索タスクの提出結果だけでなく、言語横断検索タスクの提出結果も、適合文書リストの網羅性を高めるのに貢献していることがわかり、サブタスク混合プーリングの効率性と有効性が確かめられた。

Pooling for a Large Scale Test Collection : Analysis of CLIR Task Results for the Third NTCIR Workshop

Kazuko Kuriyama* Koji Eguchi† Kazuaki Kishida‡ Noriko Kando§

Abstract. The purposes of this study is to verify the effectiveness of the subtask-mixed pooling method to construct a test collection. We carried out an experiment using the relevance assessments of NTCIR-3 and the search results submitted for the CLIR task at the third NTCIR workshop. The result is that the search results for BLIR task and MLIR task, as well as SLIR task, contributed for collecting the unique relevant documents. Hence, we verified the efficiency and effectiveness of the pooling.

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）を中心とした「情報検索システム評価用テストコレクション構築プロジェクト」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）の構築を行なっている [6]。その過程において、2001 年 9 月から 2002 年 10 月まで、評価型ワークショップ NTCIR ワークショップ 3[7] を開催し、テストコレクション 3(NTCIR-3) の構築および検索システムの評価を行なってきた。

1.2 テストコレクション

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適合文書の網羅的リスト、からなる。

適合文書の網羅的リストを作成するためには、各検索課題についてデータベース中の全文書の適合判定を行うことが必要である。しかし、数万件以上の文書を含む大規模データベースの全文書についてこれを行なうことは、時間と人的

* 白百合女子大学 Shirayuri College
† 国立情報学研究所 National Institute of Informatics
‡ 駿河台大学, 国立情報学研究所 Surugadai University
§ 国立情報学研究所 National Institute of Informatics

資源の面から考えて、非現実的である。

そのため、大規模テストコレクションの適合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、それを人間の判定者が検索課題に適合か不適合かを判定して、適合文書のリストを作成する、プーリング法が一般的に採用されている。これは、異なる検索手法を用いた検索システムは異なる適合文書を探ということが知られているからである [3]。

TREC(Text REtrieval Conference)[8] では、1992 年から毎年、多くの研究者から検索結果を収集してプーリングをすることによって、大規模テストコレクションの効率的な構築を実現している。

1.3 目的

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) 適合文書リストの網羅性:
プーリングによる適合文書収集では、プールに入れられなかった文書は不適合文書であるものと仮定される。そのため、適合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) 適合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、適合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) 適合判定の無矛盾性:
適合判定が複数の判定者によって行なわれるとき、判定者間の判定にはゆれがある。そのゆれによって、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 および NTCIR-2 構築の過程において、上記の点について、以下のような実験と考察を行なった [4] [5]。

(1) については、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適合文書リストの網羅性を高めることができた。

(2) に関しては、プーリングによって作成した数種類の適合文書リストを用いて評価を行ない、プーリングよる適合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めた。また、プーリングに含まれないシステム（ワークショップに参加していないシステム）についてプーリングで作成したテストコレクションを使用できるかどうか（テストコレクションの再利用性）を調べるために、複数のチームに関して、それぞれ、そのチームの提出結果を全く入れないプールを作成し、そのチームの提出結果が入っていないプールを適合文書リストとして用いて、そのチームの提出結果を評価した。その結果、相対的なシステム評価にはほとんど影響がなく、プーリングに参加していないシステムの評価にも、プーリング法で作成したテストコレクションを利用することができることがわかった。

(3) については、異なる 2 人の適合判定者による判定結果と最終判定結果という 3 つの異なる適合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、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を確かめた。

NTCIR ワークショップ 1 および 2 の日本語・英語検索タスクで使用した文書は、日英の対訳データを多く含み、部分的には、ほぼ対訳コーパスとなっていた。NTCIR ワークショップ 3 の言語横断検索タスクの使用文書は、同年の同主題の文書を含むが、対訳ではなく、コンパラブルコーパスになっている。

NTCIR ワークショップにおいては、3 言語以上の多言語コーパスを使用したプーリングによる適合文書リストの作成は初めての試みであったため、適合文書リストの網羅性と各 run への公平性を考慮して、サブタスク、および、検索課題と検索対象言語の組合せを区別せずに、全サブタスクの全提出結果から上位一定数をプー

ルした（サブタスク混合プーリングと呼ぶ）。

本稿では、コンパラブルコーパスに対する、網羅的で効率的なプーリング手法の確立を目的として、NTCIR-1 および NTCIR-2 構築の経験を踏まえて、サブタスク混合プーリングによる NTCIR-3 の適合文書リストの作成が有効であったかどうかを、プーリング実験を行なって考察する。

2 NTCIR-3 言語横断検索タスク

2.1 サブタスクと検索対象文書

本項以下では、NTCIR ワークショップ 3[7] の「言語横断検索タスク (Cross-Lingual IR Task)」を「CLIR タスク」と略記する。CLIR タスクには 3 つのサブタスクがある。

- 単言語検索 (SLIR): ある言語で書かれた検索課題を用いて、検索課題を同じ単言語の文書セットを検索する。
- 2 言語検索 (BLIR): ある言語で書かれた検索課題を用いて、検索課題とは異なる単言語の文書セットを検索する。
- 多言語検索 (MLIR): ある言語で書かれた検索課題を用いて、複数言語の文書セットを検索する。

各サブタスクの検索課題 (Topic) の言語と検索対象文書 (Doc) の課題の組合せ、および、各文書セットの文書数は表 1 の通りである。

表 1: Topic Sets and Document Sets

Subtask	Topic → Doc	Topic → Doc
SLIR	C → C	J → J
	E → E	K → K
BLIR	C → J	C → K
	E → J	E → K
	K → J	J → K
	J → C	K → C
	E → C	
MLIR	C → CJE	C → CJ
	E → CJE	E → CJ
	J → CJE	J → CJ
	K → CJE	K → CJ
	C → JE	C → CE
	E → JE	E → CE
	J → JE	J → CE
	K → JE	K → CE

C: Chinese, E: English, J: Japanese, K: Korean

表 2: Number of Documents

Language	Document Set	Number
Chinese	CIRB011	132,173
	CIRB020	249,508
English	Mainichi Daily News	12,723
	EIRB010	10,204
Japanese	毎日新聞	220,078
Korean	Korea Economic Daily	66,146

中国語文書セット (C)、英語文書セット (E)、日本語文書セット (J) の発行年は 1998 年〜1999 年、韓国語文書セット (K) の発行年は 1994 年である。検索課題もそれに合わせて、1998-1999 年用 50 課題と 1994 年用 30 課題の 2 種類がある。1998-1999 年用検索課題セットは C, J, E に対して使用し、1994 年用検索課題セットは K に対して使用する。ただし、各検索対象文書セットについて適合文書が 2 件以下の検索課題については、評価に使用する正式な適合文書リストからははずしているため、検索対象文書セットによって使用できる検索課題数が異なっている。中国語文書セット (C) に使用できる検索課題は 42 件、日本語文書セット (J) に対して使用できる検索課題は中国語文書セットとは異なる 42 件である。英語文書セット (E)、中国語・日本語文書セット (CJ)、中国語・英語文書セット (CE)、日本語・英語文書セット (JE)、中日英語文書セット (CJE) に対しては、それぞれ、順に、32, 50, 50, 46, 45, 50 件の課題が使用でき、韓国語文書セット (K) に対しては 30 課題が使用できる。各検索課題セットは、中国語、英語、日本語、韓国語のそれぞれに翻訳されている。

2.2 提出結果からのプーリング

CLIR 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、単言語の文書セットあるいは多言語の文書セットを検索し、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。

NTCIR ワークショップ 3 の CLIR タスクの fomal run には、23 チームが参加し、199run を提出した。そのうち、正式な run として、189run が評価されたが、適合文書リストの網羅性を高めるため、プーリングでは提出された 199run 全てを使用し、検索課題と検索対象言語の組合せに関係なく、どの run からも、同一の検索課題については同一の一定数の上位 X 件の文書をプールし、プールした文書を各言語ごとに分けて、適合判定を行なった。

このプーリングの過程では、各サブタスクの run を区別しないので、本稿では、サブタスク混合プーリングと呼ぶ。実際のプーリングに使用した run 数とチーム数を検索課題 (Topic) と検索対象文書 (Doc) の組合せごとに表 3 に示す.

サブタスク混合プーリングでは、1 つの検索課題に対してプールされた、言語ごとに文書数の合計が 2000 件程度になるように、検索課題ごとに各 run からプールされる文書数 X を決定した。中国語、英語、日本語の各文書セットを検索対象とする run では、X を 80,90,100 のいずれか、韓国語文書セットを検索対象とする run では、X を 180,190,200 のいずれかとした。プールされた適合文書候補の適合判定は、1 つの検索課題に対する 1 つの言語の文書セットに対して、当該言語を母国語として話す判定者 1 人が行なった。ただし、英語文書セットについては、EIRB010 と Mainich Daily News とを別々の文書セットとして判定した。NTCIR-1 と 2 の作成過程から、経験的に、1 つの検索課題についての文書間の判定基準になるべく矛盾がないように適合判定を行なえる上限は 2000 件程度であると考え [4] [5]、また、全過程の中で適合判定に使用できる期間を考慮して、このようなプール数の調整を行なった。同一の検索課題については、run 間の公平性を保つため、どの run からも同じ上位 X 件をプールした。

表 3: Number of Pooled Runs and Groups

Sub task	Topic -Doc	Pooled Run	Pooled Group	Total Run	Total Group
SLIR	C-C	35	14	113	21
	E-E	28	14		
	J-J	33	14		
	K-K	17	8		
BLIR	E-C	16	6	57	14
	J-C	5	3		
	K-C	2	1		
	C-E	3	1		
	J-E	1	1		
	K-E	7	1		
	C-J	4	2		
	E-J	13	6		
	E-K	6	2		
MLIR	C-CE	3	1	29	7
	E-CE	6	2		
	C-JE	3	1		
	E-JE	1	1		
	J-JE	2	2		
	C-CJ	3	1		
	C-CJ E	4	2		
	E-CJ E	4	2		
	J-CJ E	3	1		
Total		199	23	199	23

2.3 正解判定

適合判定は、(高適合) highly-relevant (S)、適合 relevant (A)、部分的適合 partially-relevant (B)、不適合 non-relevant (C) の 4 つのレベルで、検索課題ごとに、判定者 1 人が行なった。NTCIR ワークショップ 3 では、「S」と「A」を適合、「B」と「C」を不適合とした適合文書リスト (Regid) と、「S」、「A」、「B」を適合とし、「C」を不適合とした適合文書リスト (Relax) を用いた、2 つのレベルの適合文書リストで各 run の評価を行なっているが、本稿では、適合「S」、「A」、「B」を「適合文書」とした適合文書リスト (Relax) を評価に使用する.

3 プーリング実験

NTCIR ワークショップ 3 の CLIR タスクでは、適合文書リストの網羅性を高めるために、

単言語検索タスクの run 以外からも各言語の文書をプールしたが、一般的には、プールに使用する run 数が増えると、プール中の文書数も増える。適合判定を行なう際の判定期間と人的資源には限りがあるので、判定時間と判定にかかる労力を減らすためには、プール中の文書数はできるだけ少なく、かつ、その中に含まれる適合文書はできるだけ多いことが望ましい。つまり、実際のプーリングでは、全ての run をプールするのではなく、適合文書を効率的かつ公平に集められるように、なんらかの基準でプールする run を選択することが求められている。

一般的には、検索課題と検索対象文書が同じ言語で書かれている場合の検索（単言語検索）は、検索課題と検索対象文書が異なる言語で書かれている場合の検索（言語横断検索）よりも容易であると考えられるが、検索課題の内容、言語の組合せ、検索手法などによっては、言語横断検索の方がうまく適合文書を検索できる場合もあるので、一概にどちらが容易であるということは言えない。また、プーリング法では、上位一定数をプールするため、1 つの言語の文書を検索対象とする検索（単言語検索、2 言語検索）の run のみからのプールの方が、2 言語以上の文書を検索対象とする検索の run のみからのプールよりも、1run あたりの 1 言語文書に対するプール件数は多くなり、より多くの適合文書を含むのではないかと予想される。

以上のようなことから、単言語検索の run と言語横断検索（2 言語検索、多言語検索）の run をどのように選択すれば効率的かつ網羅的なプーリングになるかを調べる必要がある。

本稿では、NTCIR ワークショップ 3 の CLIR タスクで提出された 199run のうち、日本語文書を検索対象文書セットに含む 70run を用いて、SLIR、BLIR、MLIR というサブタスクごとに、日本語文書セットについてのプーリング実験を行ない、適合文書数の比較を行なった。

3.1 SLIR の run からのプーリング

NTCIR-3 の適合文書リストを R とする。BLIR タスク、MLIR タスクの run は使用せず、SLIR タスクの run のみからプールを行なった場合、適合文書の網羅性にどのような影響があるのか調べるため、R 中の J-J タスクの 33run のみからプールした文書のリスト PS を作成した。表 4 に、検索課題ごとの R と PS の適合文書数を示す。また、表 5 に、PS にだけ含まれていて、BLIR タスクと MLIR タスクの run には含まれていない、ユニークな適合文書数を示す。さらに、プール件数ごとの適合文書リストの網羅性を見るために、J-J タスクの 33run のみから、$X = 10, \cdots, 100$ について、上位 X 件をプールし、そのプールを PSX とした。図 1 に PX と PSX の適合文書数の R に対する割合の平均を示す。

表 4 と図 1 からわかるように、検索課題に対する適合文書の多寡に関係なく、PSX に含まれる適合文書の R に対する割合（カバー率）は高くなっているが、表 4 の検索課題ごとのカバー率の平均%av3（適合文書が 100 件以上ある検索課題についての、R に対する適合文書の割合）から、特に、適合文書が多い検索課題ほどカバー率が低くなっていることがわかる。このことから、検索課題全体としては、SLIR タスクの run からのプーリングだけでも、9 割の適合文書を集めることができるが、検索課題の適合文書数が多い場合には、SLIR だけでは漏れてしまう適合文書があり、BLIR タスク、MLIR タスクの run からのプーリングも網羅性を高めるためには必要であるということがわかった。

3.2 BLIR/MLIR タスクの run からのプーリング

2.2 節で述べたように、CLIR タスクでは、検索対象文書が単言語の run でるか、多言語の run でるかにかかわらず、評価のための公平性を保つため、各 run から同一の上位 X 件をプールした。

BLIR タスクの run は、単言語文書セットを検索対象としているため、多言語文書セットを検索対象とする MLIR タスクの run よりもうまく適合文書を集められるのではないと予想される。SLIR タスクの run からのプールと比較する

ため、BLIR の、日本語文書を検索対象とした 17run のみからプールを作成し、そのプールを PB とした。また、$X = 10, \cdots, 100$ について、上位 X 件をプールし、それをを PBX とした。

MLIR タスクの run からのプーリングでは、各 run の中の各言語文書の割合にかかわらず、上位 X をプールするため、、各言語に対しては、各 run からプールされる文書数は X 件以下になる。SLIR タスクからのプール PS と上記の PB と比較をするために、検索対象文書として日本語文書セットを含む MLIR タスクの 20run からのプールを作成し、PM とした。

MLIR において、言語ごとに上位 X 件の文書をプールした場合には、より多くの適合文書を見つけることができるのではないかと期待される。MLIR タスクの run からより多くの文書をプールすることによって、適合文書リストの作成ににどのような影響があるのか調べるため、NTCIR-3 を作成したときと同じのプール件数を用いて、プールを作成し、PM' とした。また、$X = 10, \cdots, 100$ について、MLIR タスクの run から言語ごとに上位 X 件をプールし、そのプールを PM'X とした。

図 1 に PBX、PMX、PM'X の適合文書数の R に対する割合の平均を示す。表 5 に PS、PB、PM の中のユニークな適合文書数を示す。図 2 に PS、PB、PM の中の適合文書の重なりを示す。図 2 では、3 つの円をそれぞれ PS、PB、PM とし、円の中の数値は、それぞれのプールに含まれる適合文書数を表わす。1 つの円の中に含まれる数値はそのプールのみでみつかったユニークな適合文書の合計である。円の重なりは共通する適合文書の集合であり、重なっている 2 つ、あるいは 3 つのプールの中の共通の適合文書の合計を表わす。

結果として、以下のことがわかった。

(1) 表5、図 2 を見ると、SLIR タスクの run が見つけたユニークな適合文書が最も多いが、BLIR タスクと MLIR タスクの run もユニーク適合文書を見つけるのに貢献している。また、図 2 の 3 つのサブタスクの run からのプールの共通な適合文書の数は 1182 であり、全適合文書数 2538 のほぼ半分が、全タスクからのプールで共通に見つかっていることがわかる。

(2) 表 4 からわかるように、PB は適合文書全体の 77.7%、PM は 66.5%をカバーしているが、PS に比べて、適合文書の漏れが多く、PB か PM のどちらかのプールだけでは、適合文書を網羅的に集めるのには不十分であると考えられる。

(3)BLIR タスクの検索対象文書は単言語であるので、多言語文書からのプールであるため、1 言語あたりの 1 run からのプール数が少ない MLIR タスクのプール (PM) よりも、BLIR タスクの run からのプール (PB) の方が多くの適合文書を含み、run の性能によっては、SLIR タスク (PS) と同程度の適合文書を含んでいるのではないかと予想していた。しかし、PB の適合文書の割合の平均は、PS よりも 20%前後小さく、単言語検索と言語横断検索の性質の違いが現れているのではないかと思われる。

(4) 表 5 を見を見ると、適合文書数が多い検索課題（012, 018〜021, 023, 036）も多く含まれているが、適合文書数が 50 件以下である検索課題、008, 031, 031 なども含まれており、適合文書が多いときに、各タスクの run がユニークな文書を多く含むとは限らないことがわかる。

(5)MLIR タスクの各 run から言語ごとに上位 X 件をプールした場合 (PM'X)、同じサブタスクの run であってもプールされる文書数がそれぞれ異なり、適合判定をされる文書数も run によって異なるため、同一タスクの run 間の評価に不公平性が生じる可能性がある。

以上のことから、BLIR タスクと MLIR タスクの run は、ユニークな適合文書数は SLIR タスクの run に比べて多くはないものの、適合文書リストの網羅性を高めるのに貢献をしていることがわかった。また、BLIR タスクの run から、言語ごとに上位 X 件をプールした場合、より多くの適合文書が含まれている可能性があるが、(5) の点から、MLIR タスクの run について、言語ごとに分けて一定数のプールを行なうのは適切でないと考えられる。

4 まとめ

本稿では、パラレルコーパスを使用する言語横断検索のテストコレクションを構築する際に、

表 4: Number of Relevant Documents in Pools

Topic	R	PS	PB	PM	PM'
002	9	9	9	8	8
004	54	51	44	12	23
005	81	80	54	44	46
007	6	6	6	6	6
008	22	22	14	14	14
010	13	13	13	12	13
012	266	168	231	126	189
014	31	31	31	29	31
015	36	36	35	35	35
016	27	27	27	25	25
017	28	28	4	22	23
018	150	134	55	22	64
019	208	198	161	106	125
020	205	188	160	136	155
021	116	111	100	85	96
022	28	28	27	21	25
023	345	316	153	72	126
024	27	27	19	22	22
025	55	55	42	54	54
026	46	46	41	35	37
027	60	60	57	53	54
028	64	63	58	48	53
029	46	46	36	32	39
030	27	27	26	25	26
031	16	15	10	11	11
032	24	24	20	18	23
033	21	21	20	9	11
034	44	37	21	8	19
035	67	62	62	53	61
036	151	145	80	72	94
037	41	41	41	41	41
038	16	16	15	16	16
039	37	37	30	28	33
040	9	9	5	1	2
041	24	24	20	21	22
042	27	27	24	19	22
043	25	23	22	16	19
044	6	6	4	5	6
045	26	26	22	12	16
046	33	32	27	16	25
047	12	12	3	2	6
050	9	9	5	4	5
Total	2538	2336	1834	1396	1721
ave	100.0	97.0	77.7	66.5	76.5
%av1	100.0	98.9	79.3	70.8	80.0
%av2	100.0	97.4	83.8	69.5	76.9
%av3	100.0	89.0	66.1	45.9	61.6

ave:平均適合文書数
%av1:適合文書数 R<50 の検索課題についての適合文書数の割合の平均
%av2:適合文書数 50≤R<100 の検索課題についての適合文書数の割合の平均
%av3:適合文書数 100≤R の検索課題についての適合文書数の割合の平均

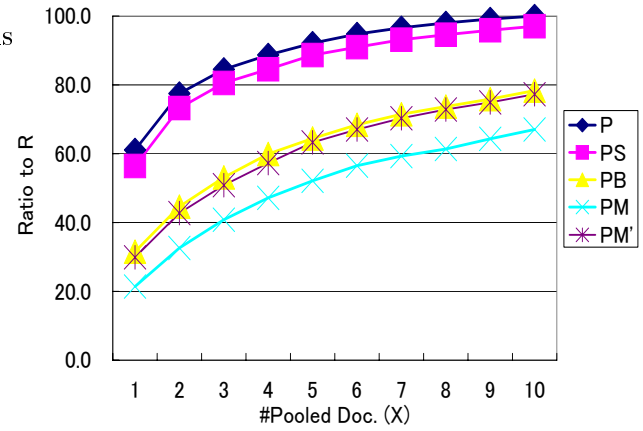

図 1: Ratio of Relevant Documents in Pools

表 5: Number of Unique Relevant Documents in Pools

Topic	PS	PB	PM
004	8	3	
005	20		
008	7		
012	28	69	4
015	1		
017	6		
018	85	13	3
019	41	6	2
020	32	11	1
021	13	2	1
022	1		
023	168	24	4
024	3		
025	1		
026	5		
027	2		
028	5		
029	8		
030	1		
031	2	1	
033	1		
034	21	3	1
035	5	2	
036	43	1	4
039	3		
040	4		
041	1		
042	1		
043	1		1
045	3		
046	4	1	
047	9		
050	2		
Total	535	136	21
%ave	18.3	2.3	0.5

%ave:各検索課題についての、ユニークな文書が適合文書数に占める割合の平均

多言語文書であるという特性を利用して、適合文書候補をどうすれば効果的かつ公平に収集することができるかという観点から、NTCIR-3 の言語横断検索タスクの提出結果を用いて、プーリング実験を行なった。

実験の結果から、NTCIR-3 については、以下

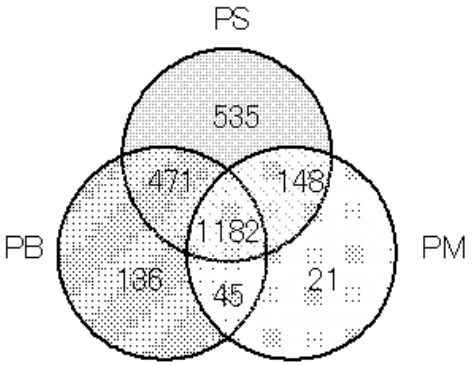

図 2: Unique Relevant Documents in Pools

のことがわかった。
(1)SLIR タスクの run のみからのプーリングでも、多くの適合文書を探すことができるが、BLIR タスク、MLIR タスクの run は、SLIR タスクの run とは異なる適合文書を見つけてくるので、適合文書リストの網羅性を高めるために、サブタスク混合プーリングを行なって、プーリングの幅を広げることは有効である。
(2)MLIR タスクの run について、各 run から上位 X をプールするのではなく、言語ごとに上位 X 件のプールを行なった場合、適合文書を効率的に集めるのには役に立つが、プール件数が run によって異なってしまい、同一タスクの run 間での相対的評価に不公平が生じるため、テストコレクションの信頼性の観点から、言語ごとに一定数のプールをを行なうのは、適切でない。

NTCIR-1 および 2 では、検索対象文書セットは単言語あるいは準対訳コーパスであったため、検索課題の言語と検索対象文書の言語の組合せは考慮せずにプーリングを行なっていた。NTCIR-3 の CLIR タスクについての分析から、BLIR、MLIR は、SLIR とは異なる適合文書を探してくることがわかり、検索対象文書だけでなく、検索課題の言語との組合せも考慮してプーリングする run を選択する必要があることがわかった。

適合文書が多い検索課題について、評価の公平性を保ちつつ、どのように網羅性を高めるかということは 1 つの課題である。NTCIR-3 の CLIR タスクでは行なわなかったが、追加プーリングや、NTCIR-1 と NTCIR-2 を構築する際に行なった、人手による追加検索による適合文書の補完作業が網羅性と公平性に与える影響についても検討が必要であると考えられる。

今後の課題として、SLIR、BLIR、MLIR に分けてプーリングを行なったときに、各サブタスクの run の評価にどのような影響があるか、また、それぞれの run の、ユニークな適合文書を見つけることへの貢献度 (unique contribution) とその評価への影響について実験を行い、考察したい。

参考文献

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

[2] Chen, K. et al., “Overview of CLIR Task at the Third NTCIR Workshop”. In Proc. NTCIR Workshop 3, Tokyo. (In printing)

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

[4] Kuriyama, K. et al., “Pooling for a Large-Scale Test Collection: An Analisis of the Search Results from the First NTCIR Workshop”. Information Retrieval, Vol.5, No.1, pp.41–59, 2002.

[5] 栗山和子ほか., “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”. 情報処理学会論文誌: データベース, Vol.43, No.SIG2 (TOD13), pp.48–59, 2002.

[6] NTCIR (NII-NACSIS Test Collection for IR Systems) Project. http://research.nii.ac.jp/ntcir/

[7] NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop 2002. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/

[8] Text REtrieval Conference (TREC). http://trec.nist.gov/

[9] Voorhees, E. The Eleventh Text Retrieval Conference (TREC 2002), NIST Special Publication SP 500-251, Maryland, U.S.A., 2002.