データベースシステム 114－3
（1998. 1. 19）

日本語情報検索システム評価用テストコレクション BMIR-J2

木谷 強 (NTT データ)　小川 泰嗣 (リコー)　石川 徹也 (図書館情報大)
木本 晴夫・中渡瀬 秀一 (NTT)　芥子 育雄 (シャープ)　豊浦 潤 (三菱電機)
福島 俊一 (NEC)　松井 くにお (富士通研)　上田 良寛 (富士ゼロックス)
酒井 哲也 (東芝)　徳永 健伸 (東工大)　鶴岡 弘 (東大地震研)　安形 輝 (慶応大)

日本語情報検索システム評価用テストコレクション BMIR-J2 は、情報処理学会データベースシステム研究会内のワーキンググループによって作成されている。BMIR-J2 は 1998 年 3 月から配布される予定であるが、これに先立ち、テスト版として BMIR-J1 が 1996 年 3 月からモニタ公開された。J1 は 50 箇所のモニタに配布され、多数の研究成果が発表されている。BMIR-J2 では、J1 に対するモニタユーザからのアンケートの回答と、作成にあたったワーキンググループメンバの経験をもとに、テストコレクションの検索対象テキスト数を大幅に増やし、検索要求と適合性判定基準も見直した。本論文では、BMIR-J2 の内容とその作成手順、および今後の課題について述べる。

BMIR-J2 - A Test Collection for Evaluation of Japanese Information Retrieval Systems

Tsuyoshi Kitani (NTT DATA)　Yasushi Ogawa (Ricoh)
Tetsuya Ishikawa (ULIS)　Haruo Kimoto, Hidekazu Nakawatase (NTT)
Ikuo Keshi (SHARP)　Jun Toyoura (Mitsubishi Electric)
Toshikazu Fukushima (NEC)　Kunio Matsui (Fujitsu Laboratories)
Yoshihiro Ueda (Fuji Xerox)　Tetsuya Sakai (Toshiba)
Takenobu Tokunaga (Tokyo Institute of Technology)
Hiroshi Tsuruoka (ERI, Univ. of Tokyo)　Teru Agata (Keio University)

BMIR-J2, a test collection for evaluation of Japanese information retrieval systems to be released in March 1998, has been developed by a working group under the Special Interest Group on Database Systems in Information Processing Society of Japan. Since March 1996, a preliminary version called BMIR-J1 has been distributed to fifty sites and used in many research projects. Based on comments from the BMIR-J1 users and our experience, we have enlarged the collection size and revised search queries and relevance assessments in BMIR-J2. In this paper, we describe BMIR-J2 and its development process, and discuss issues to be considered for improving BMIR-J2 further.

1 はじめに

近年、テキストの電子化が急速に進み、論文、図書文献をはじめとして、新聞記事、特許、オフィス文書、インターネットのホームページにいたるまで検索対象が広がってきた。蓄積されるテキストの種類と量が増加するにつれ、検索システムにおける検索精度の評価がますます重要になってきている。

検索システムの評価を同一の基準で定量的に実施するためには、共通の評価用データを使って検索結果を比較する必要がある。このような評価用テストコレクションは、欧米では古くから作成されて一般に提供されている [1][2][3]。近年は、フルテキスト形態の大規模なテストコレクションの作成が進められている [3]。特に、米国規格協会 (NIST) が 1992 年から開催している情報検索システムの評価会 TREC (Text REtrieval Conference)[4] では、分野を限定しないギガバイト規模のテストコレクションが構築・使用されている。これら欧米のテストコレクションは、TREC に代表されるような複数システム間の優劣・特徴の公正な比較、論文などで提案する手法の客観的な有効性検証、システムの設計・改良のための分析評価などに活用されている。

日本語については、長らく評価用テストコレクションがなく、情報検索システムの開発元が独自に準備したテキスト集合と基準に基づいて、方式の精度やシステム性能が評価されていた。このような背景を踏まえ、我々は 1993 年 2 月に情報処理学会データベースシステム研究会の下部組織として、「情報検索システム評価用データベース構築ワーキンググループ」(以下 WG と略記) を設立した [5][6][7]。1996 年 3 月には、日本語情報検索システム評価用テストコレクション BMIR-J1 (Benchmark for Information Retrieval Systems for Japanese Texts Ver.1) を開発し、小規模ではあるが日本語として初めてのテストコレクションを一般に公開した。BMIR-J1 は、中間成果物として WG 外からも広く意見を吸い上げるためのテスト版という位置付けで、50 箇所のモニタに公開した [8][9][10][11]。

BMIR-J1 は評価用という位置付けであったためテストコレクションの規模が小さく、WG 委員およびモニタユーザから規模の拡大を求める声が強かった。そこで、検索対象テキスト数を大幅に増やし、検索要求と適合性判定基準も見直した本格版テストコレクション BMIR-J2 を作成した。

本論文では、まず第 2 節でテスト版 BMIR-J1 の概要を紹介する。第 3 節では BMIR-J1 に対するモニタユーザからのアンケートの回答を分析する。第 4 節では、BMIR-J2 の内容を BMIR-J1 と比較しながら説明する。第 5 節で BMIR-J2 の構成と配布方法を紹介し、最後に今後の課題を述べる。

2 テスト版 BMIR-J1

BMIR-J1 は、検索対象テキスト集合、検索要求、検索要求に対する適合性判定結果 (正解テキスト集合) からなる。その内容は、[8][9][10][11] に詳しく報告されているので、本節では概要を紹介する。

2.1 検索対象テキスト集合

日本経済新聞の朝刊経済面 4 カ月分 (1993 年 9 月 1 日～12 月 31 日) [1]の 41843 件 (ただし人事異動などの記事は除く) を母集合とし、そこから無作為に 600 記事をサンプリングしたものを BMIR-J1 の検索対象テキスト集合とした。

テキスト件数については、検索システムの比較評価に符号検定を用いることなどを含めて統計的な観点から議論を行ない [5][6]、WG で作成する最終的な日本語テストコレクションでは少なくとも 6000 件規模を目標とすることにした [7]。中間成果物として WG 外からも広く意見を吸い上げるためのテスト版である BMIR-J1 では、テキスト件数を最終版の 1/10 である 600 件とした。また、テキストの形態はフルテキストとし、その内容は、テスト版の段階では単一分野、実際のビジネスシーンでの利用ニーズが高いと考えられる理由から経済面を選んだ。

各記事は、本文テキストに加えて、記事 ID、見出し文、書誌事項、日経新聞社で付与したキーワードリストなどが、SGML 形式[2]で記述されている。

2.2 検索要求と適合性判定結果

検索要求は、自然言語 1 文で記述するものとし、「～に関する記事がほしい」という形式で統一した。「～」の部分にあたる名詞句のみを列挙する形式で、60 件用意した。キーワード論理式の形式では検索要求が適切に表現できないことや、実際の利用場面やシステムへの入力のしやすさの観点から自然で簡潔な表現が好ましいことから、上記の表現形式とした。ただし、適合性判定に際し、検索意図を明確にする

[1] 日経データと情報処理学会の間で本記事データの利用に関する契約を結んだ。本新聞記事データをご提供いただいた日経データに深謝する。

[2] 毎日新聞社と NTT データが新聞記事向けに定義した DTD を用いた。本 DTD の BMIR-J1 での使用を快諾してくださった毎日新聞社メディア事業部の寒田正彦部長に深謝する。

ためには補足説明が必要になり、検索要求文(名詞句)と併せて、補足説明も付記することになった。

各検索要求に対する適合性判定結果として、正解記事IDの集合と各々の判定理由(コメント)を付加した。正解は、検索要求を主題とする記事をAランク、検索要求が主題とはなっていないが検索要求の内容を少しでも記述している記事をBランクとして2種類に分けた。これは、情報検索システムの性格や利用場面によって、Aランクのみを正解とするか、AランクとBランクの両方を正解とするかが異なってくると考えたためである。また、不正解としたものでも、正解か否かの判定が微妙だったものや、分析の参考になりそうなものについては、Cランクを付けて不正解判定の理由をコメントに残した。

2.3 作成手順

BMIR-J1は、外部へは委託せずWGメンバ(第1期の12名)で作成した[3]。テストコレクションの作成においては、適合性判定基準の妥当性と、複数人で作業を行なうことによる効率性と不均一性とのバランスが問題になる。効率を上げるため、まず検索要求と新聞記事を6つのセットに分け、各セットを2名の委員がペアとなって正解を作成した。次に、ペア同士で確認作業(クロスチェック)を実施し、個人差による品質のばらつきを低減した。さらに正解判定が難しい個別の検索要求は、WG委員全員で判断基準を確認しコレクション全体の整合性を保った。

3 BMIR-J1の評価

BMIR-J1は、情報処理学会誌にてモニタ利用を募り、1996年3月からモニタに配布を開始した。モニタ配布先の数は、検索対象テキスト集合に用いた新聞記事データの利用契約により、WGメンバの所属機関を含み50箇所に限定された[4]。モニタ公開後、アンケート(利用報告書)の提出を依頼し、これまでに32箇所から回答が集まっている(WGメンバ自身による回答も11件含まれている)。

3.1 BMIR-J1を用いた研究事例

モニタへのアンケートには、BMIR-J1を評価に利用した情報検索システムの概要を簡単に記述してもらう質問項目があった。この設問に対しては30件の回答があり、それらの検索モデルは、Boolean, Vector Space, Extended Boolean, Probablistic, Pattern Matching, Neural Networkなどにタイプ分けすることができた[5]。各モニタによりBMIR-J1を用いた多数の研究発表がされている[12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31]。

これらの研究の多くは、検索手法の精度(再現率、適合率)を計算・比較するためにBMIR-J1を用いている。しかし、そのような使われ方に限らず、正解テキスト集合をクラスタとみなした評価や、適合性フィードバックのようなユーザとのインタラクションを含めた評価などにも利用されている。

3.2 モニタユーザによるアンケート結果

アンケートでは、9つの質問事項を用意した。このうち、正解のランク分け、検索要求数の多少、およびベンチマークを拡充する方法に関し、回答を表1にまとめる。

個々の検索要求に対して、適合性判定基準や正解追加またはランク修正に関する意見が寄せられた。それらの各々についてWG内で議論したが、作業手順に不備があって生ずるような問題は見られなかった。適合性判定の解釈に関する個人差(ゆれ)の範囲におさまるものが多いように見受けられた。

全体的にはBMIR-J1の有用性を認め、今後の改良に期待する意見が多数寄せられた。中でも、テストコレクションの規模拡大を望む意見が多かった。

4 本格版BMIR-J2

BMIR-J2は、モニタユーザおよびWG委員からのアンケート結果がおおむね好評であったことをふまえ、J1の設計方針を踏襲しつつ、できるだけ多くの改良を実施する方針で開発した。主な検討項目は、対象テキストの種類と分野の拡充、コレクション規模の拡大、および正解作成手順であった。

4.1 検索対象テキストの種類と分野の拡充

第3.2節に示したアンケート結果から、新聞記事の他に拡充する対象テキストとして、特許明細書と論文・抄録が希望されていることがわかった。また、

[3] BMIR-J1の作成には、増永 良文(図書館情報大)、田中 智博(NTT)、宮内 忠信(富士ゼロックス)、三池 誠司(東芝)も参加した。

[4] J1の配布は終了したが、日経データとの契約が解消されない限り、J1モニタユーザはJ1を継続して使用することができる。

[5] 設問では「検索要求の処理方法」「検索対象文書の処理方法」「検索手法」という3点に関して自由記述形式で回答を求めた。各モニタの側で該当分類タイプを選択してもらうのではなく、自由記述の内容を筆者達が解釈して分類したものである。

表 1: アンケート集計の一部

Q2: 正解のランク分け		
A/B/C の 3 段階でよい	21	(15)
A/B の 2 段階でよい	5	(3)
1 段階でよい	2	(1)
可能な限り多段階がよい	1	(0)
その他	2	(1)
合計	31	(20)

Q3: 検索要求数の多少		
現状では少ない	16	(8)
現状で十分	10	(7)
現状では多い	1	(1)
その他	1	(1)
合計	28	(17)

Q8-1: 拡充してほしいテキストの種類		
特許明細書	14	(11)
論文・抄録	15	(5)
雑誌	3	(2)
百科辞典	3	(2)
マニュアル	3	(3)
WWW ページ	3	(1)
調査記録・修正履歴	2	(2)
判例文	2	(2)
外国語・多言語	2	(2)
合計	47	(30)

Q8-2: 拡充してほしいテキストの分野		
技術・コンピュータ	16	(9)
政治・国際	4	(3)
法律	2	(2)
社会・文化	2	(2)
趣味・家庭	2	(1)
広告・新製品紹介	2	(1)
論説・解説記事	1	(1)
雑多な分野	5	(2)
限定した分野	1	(0)
合計	35	(21)

カウント値は WG メンバを含めたアンケート回答者の合計。
括弧内のカウント値は WG 外のモニタ回答者数。
Q8 の方は複数項目の重複回答あり。

分野については、技術・コンピュータの要望が強かった。WG 委員からも、論文・抄録は実際の業務に近く、また技術・コンピュータ分野であれば正解が作成しやすいという意見があった。そこで、商用の文献検索サービスを提供している組織からデータを入手することを試みたがライセンスにはいたらず、対象テキストの種類を拡充することは断念した。新聞記事については、有償ではあるが研究目的に限り利用が認められている毎日新聞社の CD-ROM データを採用することとした。記事の分野には、J1 で対象とした経済の他に工学関係も含めることとした。

4.2 検索要求文の修正

検索要求の数に関しては BMIR-J1 は少ないという意見があったが、適合性判定の作業量を考慮して J1 と同様の 60 個とした。

J1 は日本経済新聞を検索対象としているため、一般紙である毎日新聞は記事の傾向が異なると予想された。そこで、J1 の検索要求文のうち、J2 では正解が無いか多過ぎるもの、および正解の判断が難しいものは新しく作成した基本機能に近い検索要求文に置き換えた。これは、検索要求が全体的に難しすぎるというユーザからのコメントを反映することにもなった。J1 の検索要求文 60 個のうち、J1 と同様または補足説明文を含めて若干の修正がなされたものは 49 個、J1 のものが削除され、代わりに新規追加されたものは 11 個である (1997 年 11 月 17 日現在)。J2 の検索要求文 60 個は以下のとおりである。番号が抜けている要求文は J1 には存在したが J2 では削除されたものであり、No.61 以降が J2 で新規に追加されたものである。ただし、番号が J1 と同じ検索要求文であっても、J2 では修正されているものがある。

0001:「菓子メーカー」
0002:「国内航空大手３社」
0003:「任天堂またはセガ」
0004:「農薬」
0005:「飲料品」
0006:「液晶」
0007:「ビデオデッキ」
0008:「携帯電話またはパーソナルハンディホン」
0010:「減税」
0012:「３期以上連続の減益企業」
0013:「千人以上の人員削減を計画している企業」
0014:「中国にある資本金五億円以上の合弁企業」
0015:「１ドル＝１００円を超える円高」
0016:「半導体製品の生産」
0017:「電話料金の値下げ」
0018:「所得税の減税」
0019:「コンピュータメーカーまたはコンピュータ部門の人員削減」

0020:「非製造業による現地法人」
0021:「政党に対する献金」
0022:「製販一体化」
0023:「円高による物価の低下」
0024:「冷夏の被害」
0025:「メーカーの減益に対する対策」
0026:「株価動向」
0027:「コンピュータ製品の市場動向」
0028:「日本製品の対米輸出量の実績」
0030:「業績悪化を原因とする企業の合併の事例」
0032:「銀行の経営計画」
0033:「リエンジニアリングかリストラの定義」
0034:「多角化事業の低迷」
0035:「異業種会社間の共同経営」
0037:「電機メーカーの中国への投資」
0038:「外国企業の日本への進出」
0039:「権限の役員への委譲」
0040:「管理部門の統廃合と営業部門の強化を行なう会社」
0041:「東南アジア諸国から日本への輸出」
0044:「材料・設備の現地調達」
0046:「経営陣刷新」
0047:「女性の雇用問題」
0048:「企業の社会貢献」
0050:「第3次産業のサービス向上」
0051:「逆輸入を行なう日本企業」
0052:「安売りを行なう流通業者」
0053:「トップの不況対策に関する発言」
0054:「業績不振の責任を取った経営者」
0055:「企業による配下企業の再編成」
0056:「円高対策のためのメーカーの海外進出」
0057:「行政機関が関係する不況対策」
0058:「不況におけるディスカウンターの台頭」
0061:「衛星放送」
0062:「賃貸住宅」
0063:「映画」
0065:「核兵器」
0066:「国連軍派遣」
0067:「ソフトウェア」
0068:「教育産業」
0071:「電気通信に関する規制緩和」
0073:「マンションの販売」
0074:「地価の下落」
0075:「高速道路の建設」

下記の例は、BMIR-J2の検索要求のNo.17である。最初の1行が「電話料金の値下げ」に関する記事が欲しいという検索要求文で、後半の2行が補足説明文である。

0017-1:Q:F=oxoxx:「電話料金の値下げ」
0017-1:Q:N-1:値下げの対象が電話料金であることを表
0017-1:Q:N-2:わす表現が含まれていれば正解とする。

各検索要求文には、その検索要求を正確に解釈するために情報検索システムに求められるファンクションの種類も付与されている。上記の例の先頭行に含まれる "F="以下の5つのo/xが順に、F1:基本機能(キーワード照合、シソーラス展開)、F2:数値レンジ機能、F3:構文解析機能、F4:内容解析機能(言語知識利用)、F5:知識処理機能(世界知識利用)という5つのファンクションに対応している。

4.3 テキスト集合の拡大と正解作成手順

テストコレクションの信頼性を上げるためには、検索対象となるテキスト集合の拡大は不可欠である。しかしテキスト数が増加すると、正解を作成する際にBMIR-J1で実施した全件チェックが困難となるため、テキスト規模と正解作成手順をめぐりWG内で活発な議論がなされた。

テキスト数については、テストコレクションの信頼性を向上させるためには、統計的に6000記事程度は必要であると計算されていた。BMIR-J2では、1994年版毎日新聞CD-ROMの記事に対して新情報処理開発機構(RWCP)が付与した国際十進分類(UDC)を参照し、経済および工学、工業技術一般に分類される記事を選択した。このうち内容が重複する記事を除くことで、最終的に5080件を選んだ。

正解の作成にあたっては、5080件の記事を全件チェックすることが理想的であるが、現実には作業量的に不可能である。そこで質問要求ごとに、チェック対象とする記事を絞り込む方法を採用した。プーリング方式は、複数の検索システムが出力した正解の和集合を正解候補とする手法である。これによりチェック対象を絞り込むことができる。今回は、外注会社が一次的な絞り込みを実施し、WG委員が各自の検索システムを使用して正解候補を加えることでチェック対象とする正解候補を作成した。この方法では、全ての正解をもらさず正解集合に含めることは保証できないが、作業量を考慮すると現実にはやむを得ない選択であった。WG委員による絞り込みの一例として、No.41「東南アジア諸国から日本への輸出」に使われた検索式を以下に示す。

(東南アジア OR ベトナム OR ラオス OR カンボジア OR
タイ ORミヤンマー OR ビルマ OR フィリピン OR
ブルネイ OR マレーシア OR シンガポール OR
インドネシア) AND (産米 OR 輸入 OR 輸出)

次に、絞り込んだ記事に対し全件をチェックした。適合性判定作業は、以下の3ステップで実施した。

1. 新情報処理開発機構の費用で適合性判定作業を外注した。外注先には正解の判定基準を説明し、補足説明を含む検索要求文と新聞記事を渡した。外注先では一次的な絞り込みを実施後、正解候補記事に対して正解を判定した。

2. 1の結果に対し、WG委員が担当分の検索要求に対する正解判定を見直した。正解にもれが多い検索要求については、WG委員が正解の記事候補を追加し、適合性判定基準を補足説明した上で、再度、外注会社に正解判定を依頼した。

3. 2の結果に対し、他委員の担当分をクロスチェックした。担当委員は指摘内容を分析し、正解判定を見直した。クロスチェックは主にA,Bランクの記事に対して実施した。

下記の例は、第4.2節で例にあげた検索要求No.17に対応する適合性判定結果の記述内容(一部のみ)である。検索要求番号(この例では0017-1)と記号'R'に続く8桁の数字列が記事IDであり、その次のA/B/Cがランクを表わしている[6]。

0017-1:R:00007460:A:「自動車・携帯電話」「各種料金」「大幅な値下げ」
0017-1:R:00183620:B:「値下げされた」「国際電話料金」,主題は減税効果の相殺
0017-1:R:00849190:C:PHSの低廉で多様な料金(値下げではない)

外注先からの作業結果の見直しと委員間のクロスチェックで問題となったのは、AランクとBランクの切り分けである。たとえば、No.3「任天堂またはセガ」という検索要求に対し、任天堂製のゲームの内容を説明している記事について、任天堂とゲームは深く結びついていると考えれば主題としてAランクとするが、任天堂という会社について記述していないと判断すればBランクとなる。結局、記事の主題か否かは個人の価値判断によるところが大きく、WGでの議論でも統一的な判断ができない場合は、最終的に担当したWG委員が決定した。

4.4 欧米のテストコレクションとの比較

表2に、欧米の古典的なテストコレクション8種と欧米の最近のテストコレクション8種に関するHarmanの比較表[3]を引用し、それにBMIR-J1とJ2の諸元を並べて示した。

この比較により、BMIR-J1は検索対象テキスト集合がかなり小規模であることがわかる。BMIR-J2についても、最近のテストコレクションと比較すると検索対象テキスト集合は小さめであるが、質問要求文は平均的な数といえる。BMIR-J1およびJ2の検索要求については、補足説明まで含めると詳細に記述されているといえる。

[6]BMIR-J1のC判定には全てコメントがついているが、J2にはコメントが付いていないものがある。

BMIR-J1とJ2を比較すると、J2は基本機能に近い質問要求が増えたため、平均検索要求長の名詞句部分は短くなっている。しかし、検索要求をより明確にしたため、補足説明を含めると平均検索要求長は長くなっていることがわかる。また、平均正解件数はテキスト集合の規模拡大にともない、J2ではJ1の2倍近くになっている。

表2には表われていないが、ファンクション分類した検索要求を提供しているのはBMIRのみである。表2において正解判定のランク分けは、BMIRがAとBの2ランクであるほか、OSHUMEDも2ランク、Cystic Fibrosisは6ランクに分けて提供している。その他のものはランク分けをしていない。

5 BMIR-J2の構成と配布

BMIR-J2の配布セットの構成を以下に示す。

- BMIR-J2の説明書
- 検索対象テキスト集合(新聞記事ID5080件)
- 検索要求(60個)
- 検索要求に対する適合性判定結果(正解テキスト集合)
- 情報処理学会との利用覚書
- 新情報処理開発機構(RWCP)との利用覚書

BMIR-J2のテストセットは、配布側の事務的な負担を軽減するため、J1とは違い検索対象テキスト集合を含まない。したがって、利用者は毎日新聞CD-ROM'94データ集(CD-毎日新聞94版)を別途購入する必要がある[7]。なお、BMIR-J2を使用するためには、利用覚書を情報処理学会と新情報処理開発機構の両方に提出する必要がある。入手方法の詳細は、情報処理学会誌2月号の会告欄を参照していただきたい。

6 おわりに

日本語情報検索システム評価用のテストコレクションであるBMIR-J2について、その内容と作成手順を述べた。設計思想はテスト版であるBMIR-J1を踏襲したが、検索対象テキスト数をBMIR-J1の600件から大幅に増やして5080件とし、テストコレクションとしての信頼性を高めた。

[7]毎日新聞CD-ROMの入手方法については、http://cactus.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.htmlを参照のこと。

表 2: テストコレクションの比較

コレクション名	テキスト件数	平均テキスト長 (*1)	検索要求数	平均検索要求長 (*1)	平均正解件数
Cranfield	1398	53.1 語	225	9.2 語	7.2
ADI	82	27.1 語	35	14.6 語	9.5
MEDLARS	1033	51.6 語	30	10.1 語	23.2
TIME	423	570 語	24	16.0 語	8.7
CACM	3204	24.5 語	64	10.8 語	15.3
CISI	1460	46.5 語	112	28.3 語	49.8
NPL	11429	20.0 語	100	7.2 語	22.4
INSPEC	12684	32.5 語	84	15.6 語	33.0
OSHUMED	348566	～250 語	101	～10 語	17/19.4 (*2)
Cystic Fibrosis	1239	49.7 語	100	6.8 語	6.4～31.9 (*3)
FSupp	11953	1823 語	44	17 語	35
Fed	410883	1235 語	44	17 語	56
TREC-1	741856	444.4 語	50	83 語	277
TREC-2	741856	444.4 語	50	105 語	210
TREC-3	741856	444.4 語	50	60 語	196
TREC-4	567529	842.0 語	50	10 語	130
BMIR-J1	600	733.8 字 (*6)	60	10.9 字 /94.5 字 (*5)	5.5/10.1 (*4)
BMIR-J2(*7)	5080	621.8 字 (*6)	60	9.8 字 /102.2 字 (*5)	10.7/27.8 (*4)

最初の 8 つは欧米の古典的なテストコレクション (TIME のみフルテキストを含むが、他は抄録)。
次の 8 つは欧米の最近のテストコレクション (OSHUMED 以外はフルテキスト)。
これら欧米の 16 コレクションの情報は文献 [3] からの引用である。
(*1) 欧米のテストコレクションでは語数でカウントしているが、有効語数は stoplist や stemming によって変化するので、ここに示した語数は比較のための目安でしかない。
(*2) 正解が 2 ランクある。
(*3) 正解が 6 ランクある。
(*4) 正解が 2 ランクある。(数値は A/A+B の件数)
(*5) 検索要求の名詞句部分のみの長さと、補足説明まで含めた長さ。
(*6) 見出し文タグと文章タグの領域を合わせた長さ。ただし、タグ自身は長さに含まない。
(*7) BMIR-J2 は 1997 年 11 月 18 日現在のもの。配布データは多少異なる可能性がある。

これまでに、BMIR-J1 のモニタユーザによる研究発表が活発に行なわれ、情報検索システムの比較評価や提案手法の有効性検証に使用されている。初めての日本語テストコレクションとしての意義・役割が理解され、情報検索分野の研究発展のために有効に活用されている。本格版である BMIR-J2 は、テストコレクションとしての信頼性が高く配布制限もないことから、広く普及し利用されることを期待する。

BMIR は他のテストコレクションとは異なり、情報検索システムに求められるファンクションの観点を導入し、検索要求を分類して提供している。これは、高精度な情報検索システムの実現へ向けて、自然言語処理技術と知識処理技術を活用した取り組みが期待される課題を、具体的な検索要求文の形で示したものである。

BMIR-J2 は、情報処理学会データベースシステム研究会の下部組織として設立された「情報検索システム評価用データベース構築ワーキンググループ」の第 2 期活動 (1996 年 4 月～ 1998 年 3 月) の成果物である。本ワーキンググループは 1998 年 3 月で活動を終了し、その後は同研究会の第三分科会が活動を引き継ぐ予定である。 BMIR-J1 の作成には約 3 年、J2 の作成には 2 年の年月を要した。ボランティアとしての WG 委員の活動には時間的制約があり、また外部に作業を委託するにも費用確保の問題があった。今後、利用者からの J2 に対するコメントや、一層の規模拡大と異なる種類の検索テキストの使用など、新たなテストコレクション作成に対する要望が出てくるものと予想される。このようなフィードバックに対応する体制の確立は今後の検討課題である。

最後に、適合性判定の議論および実作業に参加していただいた学術情報センターの神門典子氏に感謝

する。また、BMIR-J2の適合性判定結果の作成および配布作業に対しご支援いただいている新情報処理開発機構に深謝する。さらに、情報検索システム評価用データベース構築ワーキンググループの活動をご支援いただいている情報処理学会データベースシステム研究会の田中克己主査に深謝する。

参考文献

[1] E. A. Fox: "Characterization of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts", Technical Report 83-561, Cornel University, 1983.

[2] 木本 晴夫: "自動索引システムと情報検索システムの評価用共通データベースの事例", 情処研報 DBS-90-10, pp. 83-92, 1992.

[3] D. Harman (Moderator): Panel: "Building and Using Text Collections", Proceedings of ACM SIGIR'96, pp. 335-337, 1996.

[4] D. Harman: "Overview of the Fourth Text REtrieval Conference (TREC-4)", National Institute of Standards and Technology, pp. 1-23, 1995.

[5] 木本 晴夫ほか: "情報検索システム評価用データベースの構築の提案", 情処研報 FI-32-1, pp. 1-8, 1993.

[6] 石川 徹也ほか: "情報検索システムの評価のためのベンチマークデータベースの構築", 情報処理学会 ADBS'93, pp. 217-226, 1993.

[7] 小川 泰嗣ほか: "日本語情報検索システムのためのベンチマークの構築", 情処研報 DBS-100-16, pp. 145-152, 1994.

[8] 芥子 育雄ほか: "情報検索システム評価用ベンチマーク Ver.1.0 (BMIR-J1) について", 情処研報 DBS-106-19, pp. 139-145, 1996.

[9] K. Matsui, et al.: "Test Collection for Japanese Information Retrieval Systems from the viewpoint of evaluating system functions", Proceedings of IROL'96, pp. 42-47, 1996.

[10] 福島 俊一ほか: "日本語情報検索システム評価用テストコレクション BMIR-J1", 自然言語処理シンポジウム「大規模資源と自然言語処理」, 1996.

[11] 木本 晴夫ほか: "日本語情報検索システム評価用テストコレクションの構築", 1998年情報学シンポジウム, 1998.

[12] Y. Ogawa: "Effective and Efficient Document Ranking without Using a Large Lexicon", Proceedings of VLDB'96, pp. 192-202, 1996.

[13] 高木 徹, 木谷 強: "単語出現共起関係を用いた文書重要度付与の検討", 情処研報 FI-41-8, pp. 61-68, 1996.

[14] 伊藤 史朗, 大谷 紀子, 柴田 省吾, 上田 隆也, 池田 裕治: "フロー情報収集・活用のための知的検索システム Fit (2) 処理方式", 情処53全大 2T-9, pp. 3-185-186, 1996.

[15] 大谷 紀子, 伊藤 史朗, 柴田 省吾, 上田 隆也, 池田 裕治: "フロー情報収集・活用のための知的検索システム Fit (3) 類似度判定", 情処53全大 2T-10, pp. 3-187-188, 1996.

[16] 菅井 猛, 和田 光教, 森田 幸伯: "WWW上の電子新聞に対する情報フィルタリング", 情処53全大 4T-8, pp. 3-223-224, 1996.

[17] 菅井 猛, 和田 光教: "WWW上の電子新聞に対する情報フィルタリングとその評価", 情処研報 FI-43-13, pp. 89-96, 1996.

[18] 山田 剛一, 森 辰則, 中川 裕志: "情報検索のための複合語マッチング", 情処研報 FI-43-5 (NL-115-13), pp. 91-97, 1996.

[19] 木谷 強, 高木 徹, 木原 誠, 関根 道隆: "フルテキストと抽出キーワードを利用した情報検索", 情処研報 FI-43-10 (NL-115-18), pp. 71-76, 1996.

[20] 野口 直彦, 稲葉 光昭, 野本 昌子, 菅野 祐司: "単語統計情報と言語情報とを併用した新しい文書検索のモデル", 情処研報 FI-44-5, pp. 33-40, 1996.

[21] 隅田 英一郎, 飯田 仁: "統計的な抄録法を使った情報検索", 言語処理学会第3回年次大会発表論文集, pp. 353-356, 1997.

[22] 塩見 隆一, 徳田 克巳, 青山 昇一, 柿ヶ原 康二: "シソーラスを用いた文書データの自動分類法", 情処研報 NL-117-14, pp. 99-104, 1997.

[23] 山田 剛一, 斎藤 公一, 森 辰則, 中川 裕志: "複合語マッチングによる情報検索", 言語処理学会第3回年次大会発表論文集, pp. 369-372, 1997.

[24] 山田 剛一, 斎藤 公一, 森 辰則, 中川 裕志: "複合語マッチングによる情報検索", 情処54全大 4K-3, pp. 3-27-28, 1997.

[25] 佐藤 進也, 神林 隆, 清水 奨, ポール フランシス: "広域分散検索と高再現率検索の結合について", 信学技報 DE96-77 (1997-01), pp. 19-24, 1997.

[26] 酒井 哲也, 梶浦 正浩, 住田 一男: "情報フィルタリングシステム NEAT のための検索要求文からのプロファイル生成", 情処研報 FI-47-12 (NL-121-20), pp. 83-88, 1997.

[27] 酒井 哲也, 梶浦 正浩, 三池 誠司, 佐藤 誠, 住田 一男: "ベンチマーク BMIR-J1 を用いた情報フィルタリングシステム NEAT の評価", 情処54全大 1S-11, pp. 3-301-302, 1997.

[28] 中島 浩之, 木谷 強: "単語の文書頻度を利用した決定木学習アルゴリズムによる relevance feedback の高精度化", 情処研報 FI-45-2, pp. 7-12, 1997.

[29] 高木 徹, 木谷 強, 関根 道隆, 出口 信吾: "シソーラス掲載語の重要性を考慮した文書スコアリング", 情処研報 FI-47-13, pp. 89-94, 1997.

[30] Y. Ogawa, T. Matsuda: "Overlapping statistical word indexing: A new indexing method for Japanese documents", Proceedings of 20th ACM SIGIR Conf., pp. 226-234, 1997.

[31] 小川 泰嗣, 松田 透: "ランキング文書検索におけるスコア合成法の評価", 情処研報 FI-47-14, pp. 95-100, 1997.