情報処理最前線

情報検索システムの評価プロジェクト：NTCIRワークショップ

神門 典子　国立情報学研究所

日本語を用いた情報検索，言語横断検索および関連分野の研究を促進するためには，各種手法の比較を可能にし，効果に関する分析的比較が行える研究基準データの整備が不可欠である．

この目的で学術情報センターでテストコレクションが構築され，これを用いた評価ワークショップが行われ，海外からの参加者を含め，多くの成果を得ることができた．また，同時期にIREXという新聞記事を用いた日本語情報検索と情報抽出（固有表現抽出）の評価ワークショップも開催され，成功を収めた．

本稿は，テストコレクションおよびそれを用いて共通の基盤でのシステム間相互比較と自由な雰囲気での参加者間の意見交換を意図したワークショップから得られた成果などを紹介し，テストコレクションおよび評価プロジェクトの意義，NTCIRとIREXの来年の予定，情報検索研究の今後の展望について述べる．

情報検索システムの評価：評価ワークショップと大規模テストコレクション

インターネット上のサーチエンジンの普及により，情報検索技術が多くの人にとって身近なものになるとともに，情報検索システムの研究開発に対する関心も高まっている．情報検索システムの研究開発には，各種の手法やアルゴリズム間の比較を可能にし，その効果に関する分析的比較評価を行うための研究基準データの整備が必要である．

■ テストコレクション ■

テストコレクションは，検索の質，すなわち検索がどのくらいうまくいったかを示す**検索有効性**を評価する検索実験用データセットで，(1) 検索対象となる文書群（データベース），(2) 利用者の検索要求を記述した検索課題群，(3) 各検索課題に適合する正解文書の網羅的リスト，からなる．

テストコレクションの歴史は，1960年代中頃に英国で行われたクランフィールド実験IIまでさかのぼることができる．テストコレクションの構築は，検索課題の収集や正解判定など大きな労力が必要であるが，いったん構築されたテストコレクションは独立して他の検索実験にも利用できるため，以後，欧米では，**表-1**に示すように，種々のテストコレクションが，多くの研究者が共通に利用できる**標準テストコレクション**として用いられてきた．

1992年からは，米国のDARPAの情報検索・抽出技術推進プログラムTIPSTERの一環として，毎年，情報検索の評価ワークショップTREC（Text REtrieval Conference）[1]が開催され，実用の検索システムが扱うデータベースに匹敵する大規模なテストコレクションが構築されている．これらの標準テストコレクションは，検索実験の遂行を容易にし，同一コレクションを使用したシステムの間で分析的な相互比較を可能にすることによって，情報検索研究の発展に多大な貢献をしてきた．

■ 日本における動向 ■

我が国でも，近年，標準テストコレクションの必要性が強く認識され，情報処理学会データベース研究会のワーキンググループにより，日本語テストコレクションBMIR-J2が構築され[2]，さらに，1999年の8月から9月にかけて，情報検索と情報抽出（固有表現抽出と用語抽出）に関して，

IPSJ Magazine Vol.41 No.6 June 2000

コレクション	文書の種類，対象分野など	言語	文書数	検索課題数	判定	入手先
ADI	情報学，抄録	英語	82	35	2	*1
CACM	コンピュータ科学，抄録	英語	3,200	64	2	*1
CISI	図書館学，抄録	英語	1,460	76	2	*1
CRAN	航空学，抄録（Cranfield実験IIで使用）	英語	1,400	225	2	*1
MED	医学（Medline），抄録	英語	1,033	30	2	*1
NLM	医学，抄録	英語	3,078	155	2	*1
NPL	電子工学，抄録	英語	11,429	100	2	*1
TIME	新聞記事全般，全文	英語	423	83	2	*1
OHSUMED	医学（Medline1987-1991から抽出），抄録	英語	348,566	106	3	*2
Cystic Fibrosis	医学（Cystic Fibrosisに関する文献），抄録	英語	1,239	100	7	*3
TREC-1, 2, 3	新聞記事，行政文書など各種，全文	英語	各741,856	各50	2	*4
TREC-4	新聞記事，行政文書など各種，全文	英語	567,529	50	2	*4
TREC-5	新聞記事，行政文書など各種，全文	英語	524,929	50	2	*4
TREC-6, 7	新聞記事，行政文書など各種，全文	英語	各556,077	50	2	*4
KRIST	科学技術論文の抄録	韓国語	13,515	30	2	
BMIR-J1	新聞記事（日経），全文	日本語	600	60	3	配布終了
BMIR-J2	新聞記事（毎日1994から抽出），全文	日本語	5,080	50	3	*5
IREX	新聞記事（毎日1994, 1995年），全文	日本語	211,853	36	3	*6
NTCIR-1	学会発表論文全分野，著者抄録	日，英	339,483	83（300）*8	3	*7

表-1　主なテストコレクション

*1: SMARTコレクションftp://ftp.cs.cornell.edu/pub/smart/
*2: ftp://medir.ohsu.edu/pub/ohsumed
*3: ftp://ils.unc.edu/pub/research/cfdbase
*4: 検索課題，正解判定http://trec.nist.gov/，文書はLDCで購入.
*5: http://www.ulis.ac/jp/~ishikawa/bmir参照．内容についての問合せはE-mail:bmir-users@rd.nacsis.ac.jp.
*5, *6ともに文書は各自が毎日新聞社から購入.
*6: http://cs.nyu.edu/cs/projects/proteus/irex
*7: http://www.rd.nacsis.ac.jp/~ntcadm/
*8: 現在配布済の検索課題は83件．最終目標は300件の予定.

IREX（アイレックス; Information Retrieval and Extraction）[3]とNTCIR（エンティサイル; NACSIS Test Collection for Information Retrieval）[4]という2つの評価ワークショップが，国内外からの参加者を集めて開催された.

この2つのワークショップを通じて構築されたテストコレクションなどの評価用データは，現在，ワークショップ参加者以外も研究目的での使用が可能になっている．2000年度からは両者が合同してNTCIRワークショップ2として活動を進める予定である[5].

本稿では，以下，テストコレクションと評価ワークショップについて概説し，続いて，IREXとNTCIRワークショップの概要を紹介し，最後に，今後の展望を示す.

テストコレクションと評価ワークショップ

■ 情報検索システム評価実験の発展の経緯 ■

情報検索研究における主要な評価実験と発展の経緯は，文献6）に概説した．情報検索は，1950年代から研究が盛んになり，やがて，実用としての**データベース検索サービス**と，**実験的情報検索研究**という2つの流れに分かれ，両者がまったく異なる発展過程をたどってきた.

実用サービスでは，ブール型検索と転置索引という初期に開発された検索技術を用い，データベースの大規模化，検索の高速化に対応してきた．他方，**実験的研究**は，小規模なテストコレクションを用い，ベクトル空間型，確率型などの多様な検索モデル・技法を開発してきた．いいかえれば，実用化では大量データを高速に処理する**効率（efficiency）**を重視したのに対し，実験的研究では，ノイズが少なく，精度の高い検索をする**効果（effectiveness）**を重視してきた.

1980年代になると，実用サービスでは，データベース規模が拡大するとともに，output overload, zero hitなどの，大規模データに対するブール型検索の限界が明らかになり，実用システムでも新しい効果的な検索手法の導入が模索された．しかし，当時，実験的研究では，文書数が数百～数千件程度の小規模なテストコレクションを用いて検索実験を行っており，研究成果が大規模データベースを扱う実用的環境に適応可能であるか明らかでなかった．TREC開始以降は，実用規模の文書データを用いた検索実験が可能になり，研究成果の実用システムへの技術移転が促進された.

たとえば，**適合度順出力**は1960年代に小規模データの実験で有用性が示されていたが，1992年以降，ようやく，実用データベース検索サービスやインターネット・サーチエンジンに取り入れられるようになった．国際会議などでもTiP（理論から実用へ; Theory into Practice）が関心を集めており，今後は技術移転が一層促進されると期待される．情報検索の実験的研究においても，TREC開始以降は，大

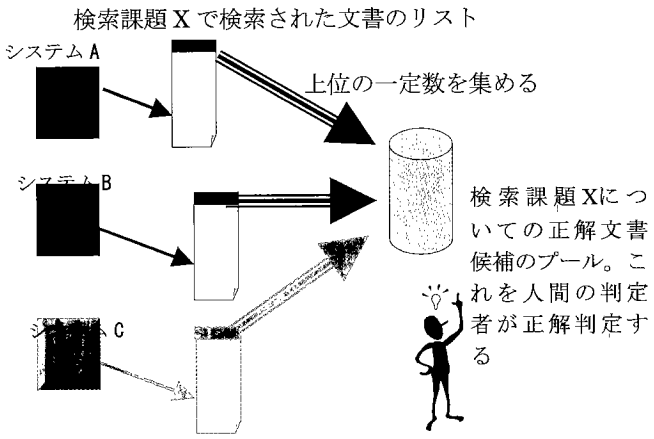


図-1　プーリングによる正解文書候補の収集

規模なテストレコレクションを用いた検索実験によって，有用性を示すことが求められるようになっている．

■ 大規模テストコレクションの構築法 ■

テストコレクションでは，各検索課題に適合する正解文書の，できるかぎり網羅的なリストが必要である．適合性の判定は，人間が1件ずつ文書を読んで内容が検索課題に適合するか判定する作業であり，1人の判定者が一貫した基準で判定できる文書数には自ずと上限があり，経験的には数千件程度までである．

したがって，対象文書数が数十万件以上となる大規模なテストコレクションでは，全数調査は不可能であり，正解文書リストは，**プーリング**によって作成する．これは，複数の異なる検索システムの同一の検索課題に対する検索結果を正解文書の候補として集め（プールし），その中の文書についてのみ，人間の判定者が1件ずつ正解判定を行う方法である．これは，Karen Spark Jonesらが提唱した大規模なテストコレクションにおいて正解文書の候補を効率よく収集する方法である[7]（図-1参照）．

TRECやNTICRなどの情報検索の評価ワークショップでは，参加者が提出する検索結果は，ランク付き，すなわち，システムが，各文書に検索課題に適合する確信度を付与し，その値の高い順に出力したリストである．各検索結果リストから，検索課題ごとに，上位一定数の文書を集めてプーリングを行う．異なる検索モデルに基づく検索システムは，異なる正解文書を検索できることが知られており，多数の検索システムが同一の課題を遂行する評価ワークショップは，プーリングの最良の機会でもある．

■ 評価ワークショップの意義と効果 ■

評価ワークショップは，複数の研究グループが，共通のデータセットを用いて，それぞれのアプローチで共通の研究タスクを遂行し，その成果を共通の評価法に基づいて分析的に比較することによって，各手法の効果を明らかにしていく研究形態である．これは，大規模な正解付きデータセットの構築，情報・技術移転の促進，特定研究課題の集中的研究，研究者が互いに知り合って気楽に意見交換できる“Who's Who環境”の生成，評価手法の標準化など，種々の成果が期待される．

たとえば，TRECは評価ワークショップの成功例の1つと考えられるが，今までTRECが取り上げてきたサブトラックには，次のようなものがある（括弧内はサブトラック名）．複数データベースの統合（Multiple Database Merging），OCR文書の検索（Confusion），英語以外の言語の検索（Chinese, Spanish），対話型検索システムの評価（Interactive），自然言語処理を用いた検索（NLP），フィルタリング（Filtering），100GBのWeb文書群の検索（Very-Large），言語横断検索（Cross Language），高精度検索（High Precision），質問回答（Q&A），検索式の収集と評価（Query）などである．これらは，特定技術の集中的研究推進や研究の動機付けを与えるものとして貢献してきた．

表-2にTRECで用いられてきた主な検索手法の傾向を示した．あるグループが提案し，有効性が示された手法は，その後，広く用いられるようになっており，TRECは，研究アイディアを交換する場として機能している．

TRECでは，毎年，使用するテストコレクションが異な

	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6
Term weighting	baseline beginning of Okapi weighting	Okapi perfects "BM25" algorithm	new SMART, INQUERY weighting algorithms	use of Okapi/SMART wt. algorithms by other groups	adaptations of Okapi/SMART algorithm in most systems
Passages	use of subdocs by PIRCS system	heavy use of passages/subdocs	decline in use of passages		use of passages in relevance feedback
Automatic query expansion		initial expansion using top X documents	heavy use expansion using top X documents	more complex expansion schemes	more groups with more sophisticated experiments
Manual query modification		manual expansion using other sources	experiments in manual editing/user-in-the-loop	extensive user-in-the-loop experiments	return to simpler editing/user-in-the-loop
Data fusion (diverse experiments)	combining manual queries	combining query mod. techniques		combining automatic queries	more complex combining of techniques
Topic analysis				initial topic "mining"	more groups/more complex topic "mining"

表-2　TRECにおける検索技術の推移（Harman, 99より）

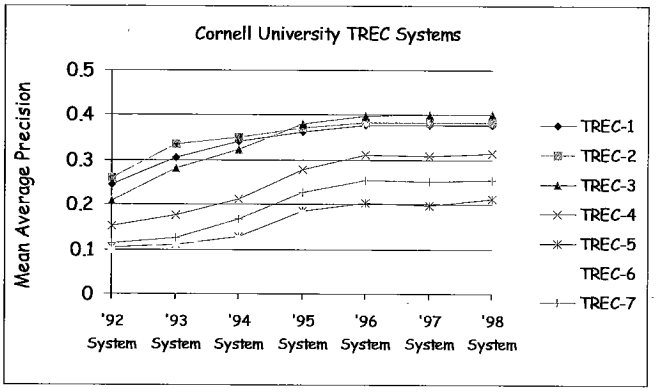


図-2　TRECにおける検索性能の向上（コーネル大学の例, Harman 99より）

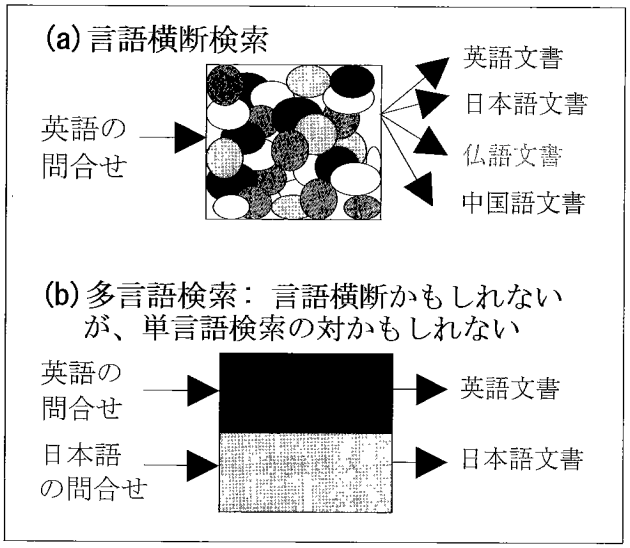


図-3　言語横断検索 vs. 多言語検索

るので，経年的な検索有効性の比較はできないが，コーネル大学の検索システムSMARTは経年的なシステムの向上を示すデータが公表されている．図-2に示すように，過去7回のTRECで使用されたテストコレクションのすべてについて，SMARTの各TRECに参加した時点のバージョンで検索した結果をみると，いずれのコレクションに対しても，次第に検索有効性が向上していることが分かる．

■ 評価手法の標準化 ■

検索有効性評価の基本となる尺度は，精度（precision）と再現率（recall）である．前者は検索された中で利用者の検索要求に適合するものの比率，後者はデータベース中の全適合文書の中で検索されたものの比率を示す．ランク付きの検索結果については，一般に以下の尺度が用いられる．

(1) 補間なしの全適合文書の平均精度（non-interpolated average precision over all relevant documents）：最も代表的な尺度．一尺度で検索の有効性を表現できる．

(2) 再現率・精度グラフ：再現率が0, 0.1, 0.2, …, 0.9, 1.0の11点での精度をグラフ化．システムのバランスを見ることができる．

(3) R精度（R-precision; 適合文書総数と同数出力した時点の精度）：主に上位ランクの検索有効性を示す．

(4) 上位ランク精度（上位5, 10, 20, 30件など一定件数の文書を出力した時点の精度）．

検索実験は通常，多数の検索課題について検索を行い，上記の評価値の全検索課題に対する平均値を求める．検索課題数は，一般に多いほどよいとされ，最低でも30件は必要と考えられる．

これらの評価値は，コーネル大学のFTPサイトから入手可能なtrec_evalプログラムによって算出できる．TRECのほか，IREX，NTCIRでも既存の評価値との一貫性という観点から情報検索系タスクでは，これらの評価値を採用した．

また，テストコレクションを用いた評価結果は，同一コレクションを用いたシステム間の相対的な比較であり，異なるコレクションから算出された評価値を直接比較することはできないということに留意する必要がある．テストコレクションはそれぞれ特徴があるので，異なるテストコレクションを用いて評価を行うとより確実な結果を得ることができる．

■ 言語横断検索（Cross-Lingual Retrieval） ■

NTCIR-1が，国際的に着目されている理由の1つは，言語横断検索研究に用いることのできるコレクションとしてである．言語横断検索とは，図-3（a）に示したように，問合せとは異なる言語で書かれた文書も一括して検索する技術で，インターネット環境下では，非常に重要である．

なお，「多言語検索（Multilingual retrieval）」は，言語横断検索を意味する場合もあるが，図-3（b）に示したように，完全対訳データベースを作成し，日本語の問合せは日本語データベース部分のみを検索するというように単言語検索の集合を意味する場合もある．したがって，言語の差を**越える，複数の言語にわたって**，という意味を明確にするために**言語横断**，Cross-Lingual，Cross-Language☆1などという用語を用いる．

ここ数年，研究が非常に盛んになり，TRECのサブトラック，NTCIRワークショップのほか，TDT（Topic Detection and Tracking）[8]，TRECの言語横断検索トラックから独立した欧州諸言語を対象とするCLEF（Cross-Language Evaluation Forum）[9]などいくつかの評価プロジェクトが計画されている．また，TRECを始めとするTIPSTERプログラムは，2000年から始まるTIDES

☆1 1996年のACM-SIGIR併設ワークショップにおいて，“Cross-”という用語を用いることが確認され，それ以降，研究コミュニティで定着している．なお，米国DARPAの研究助成プログラムでは，translingualという用語を用いる．

(Translingual Information Detection, Extraction and Summarization) に引き継がれる．範囲も欧米諸語間から，英語と日本語や中国語といった，まったく構造や起源の異なる言語間での研究へと広がっている．

なお，日本語文書の検索では，専門用語や新しい用語は，外国語の原綴やカタカナ表記の場合も多く，日本語の文書のみを扱う場合でも，これらの表記の多様性に対処するにも言語横断技術が必要である．

日本におけるテストコレクション構築と評価ワークショップ

日本初の情報検索と情報抽出・用語抽出の評価ワークショップであるIREXとNTCIRは，互いに協力関係を保ちながら活動を進め，1999年8月～9月に，合同ワークショップを挟み，成果報告会を連続開催した．概要については，それぞれの会議録収載のOverviewを参照されたい[3), 4)]．

IREX

■ 背景と目的 ■

IREXは，ニューヨーク大学の関根氏が提案，通信総合研究所の伊佐原氏と協力して，推進したプロジェクトである．自然言語処理研究者を中心として組織され，テキストを対象とした自然言語処理技術の研究に資する評価用データの構築と評価を行う．第1回は情報抽出（固有表現抽出，NE）と情報検索を取り上げた．以下の効果が期待される[3)]．

- 問題点共有とそれに基づく分野の飛躍的発展
- 情報検索，情報抽出研究の認知，宣伝，推進
- 研究者のすそ野の拡大
- データの蓄積，テキストデータベースの拡大
- 長期的なプロジェクトへの緒

■ タスク ■

以下の2種類のタスクを設定した．

固有表現抽出（NE）：新聞記事から，組織名，人名，地名，時間表現などの自動的な抽出を行う．

情報検索（IR）：文書から検索課題に適合する文書を検索する随時検索（ad hoc IR）．

■コレクション ■

IREXで使用したテストコレクションは，次の通りである．

(1) **文書**：毎日新聞の1994年と1995年の2年分の記事（211,853記事）を使用した．参加者は各自で毎日新聞社からCD-ROMを購入し，研究目的使用の許諾を得る．NEタスクでは，このうち，トピックを限定しない72記事とトピックを限定した20記事を使用した．

(2) **検索課題**：内容はBMIRに準じ，形式的には，TREC, NTCIRと同等のタグを用いている．検索課題数は，予備試験6個，本試験30個であった．

(3) **正解判定**：BMIRと同様の3段階の正解判定（A: 記事の主題が検索課題に関連，B: 主題ではないが記事の一部が関連，C: 関連しない）．判定は，判定者が行い，2名のクロスチェックをした．

■ 日程・参加者 ■

1998年11月	：予備試験（IR，NE）
1999年2月末日	：申込み締切
4月5日	：IR本試験用検索課題配布
4月12日	：IR検索結果提出
5月13日	：NE本試験用課題配布
5月17日	：NE結果提出
9月2～3日	：成果報告会

最終的な参加者は情報検索タスク14，固有表現抽出が15，計25グループであった．情報検索は国内からの参加グループのみで，企業からの参加は6グループであった．固有表現抽出の定義作成，正解作成，情報検索の正解作成，データ提供，実行委員などに協力した団体も多く，それらを合わせると総計45の組織が何らかの形でIREXにかかわったことになる．

また，IREXでは，特にNEタスクの定義と評価について，メーリングリストを通じて，非常に熱心な議論が行われた．このように，関心のある研究者が評価法について，オープンで集中的な議論をする場を提供することも，このような評価ワークショップの1つの効果といえるだろう．

NTCIR

学術情報センター研究開発部では，(1) 大規模な日本語テストコレクションを構築することによって，日本語情報検索，言語横断検索研究を促進し，基礎的な研究成果を蓄積，(2) 大規模テストコレクションの効率的構築法検討，(3) 対話型システムの評価法の検討などの目的で，1997年からNTCIR/NACSIS Corpusプロジェクトを開始し，テストコレクションの構築を進めてきた．

その過程で，テストコレクション1（NTCIR-1）を用いてNTCIRワークショップを開催した．その目的は，(a) 参加者からの正解文書候補とフィードバックの収集，(b)

<REC>
<ACCN>gakkai-0000011144</ACCN>
<TITL TYPE="kanji">電子原稿・電子出版・電子図書館-「SGML実験誌」の作成実験を通して</TITL>
<TITE TYPE="alpha">Electronic manuscripts, electronic publishing and electronic library</TITE>
<AUPK TYPE="kanji">根岸 正光</AUPK>
<AUPE TYPE="alpha">Negishi, Masamitsu</AUPE>
<CONF TYPE="kanji">研究発表会（情報学基礎）</CONF>
<CNFE TYPE="alpha">The Special Interest Group Notes of IPSJ</CNFE>
<CNFD>1991.11.19</CNFD>
<ABST TYPE="kanji"><ABST.P>電子出版というキーワードを中心に，文献の執筆，編集，印刷，流通の過程の電子化について，その現状を整理して今後の動向を検討する．とくに，電子出版に関する国際規格であるSGML（Standard Generalized Markup Language）に対するわが国での動きに注目し，学術情報センターにおける「SGML実験誌」およびその全文CD-ROM版の作成実験を通じて得られた見知を報告する．また電子図書館について，その諸形態を展望する．出版文化に依拠するこの種の社会システムの場合，技術的な問題というのは，その技術の社会的な受容・浸透の問題であり，この観点から標準化の重要性を論じる．</ABST.P></ABST>
<ABSE TYPE="alpha"><ABST.P>Current situation on electronic processing in preparation, editing, printing and distribution of documents is summarized and its future trend is discussed, with focus on the concept: "Electronic publishing". Movements in the country concerning an international standard on electronic publishing, SGML (Standard Generalized Markup Language), are assumed to be important, and the results from an experiment at NACSIS to publish "SGML Experimental Journal" and to make its full-text CD-ROM version are reported. Various forms of "Electronic library" are also investigated. The author puts emphasis on standardization, as technological problems for those social systems based on cultural settings of publication of the country, are the problems of acceptance and penetration of the technology in the society.</ABSE.P></ABSE>
<KYWD TYPE="kanji">電子出版 // 電子図書館 // 電子原稿 // SGML // 学術情報センター // 全文データベース</KYWD>
<KYWE TYPE="alpha">Electronic publishing // Electronic library // Electronic manuscripts // SGML // NACSIS // Full text databases</KYWE>
<SOCN TYPE="kanji">情報処理学会</SOCN>
<SOCE TYPE="alpha">Information Processing Society of Japan</SOCE>
</REC>

図-4　NTCIR-1の文書レコードの例

各種技法の効果を共通の基盤で比較し，(c) 意見交換をする場を提供することである．主な関心事は，数値によるシステムの順位付けではなく，どのような技術を用いるとどのような効果があるかというシステム間の比較と特徴付けである．

■ タスク： ■

第1回NTCIRタスクは以下の3種である．参加者は1つ以上のタスクを遂行した．

随時検索タスク（ad hoc IR task）：特定のデータベースに対して，新しい検索課題の検索を行い，その検索性能を調べる．

言語横断検索タスク（cross-lingual IR task）：日本語の検索課題を用いて，英語の文献を検索する．

用語抽出・役割分析タスク（automatic term recognition and role analysis task）：標題および抄録から用語を抽出し，抄録の主要論述における「対象」「手法」「主操作」を表す用語を識別する．

NTCIRワークショップの随時検索タスクは，日本語と英語が混在したデータベースを使用しており，これも言語横断検索の一種である．

<検索課題 q=0005>
<タイトル>
特徴次元リダクション
</タイトル>
<検索要求>
クラスタリングにおける特徴次元リダクション
</検索要求>
<検索要求説明>
オブジェクトのクラスタリングを行うとき，オブジェクトを特徴ベクトルで表現することが望まれる．アプリケーションによっては，オブジェクトの次元は数千，数万となることがある．このような場合，事前に次元を落とすことが必要になる．正解文書は，特徴次元リダクションの方法について，理論面から，または実験によって，提案，比較などを行っているもの．画像処理などの実験の操作の一部として特徴次元リダクションを用いているだけでは要求を満たさない．
</検索要求説明>
<概念>
特徴選択，主成分分析，情報の粒度，幾何クラスタリング
</概念>
<分野>
1．電子・情報・制御
</分野>
</検索課題>

図-5　NTCIR-1の検索課題の例

■ 参加者： ■

6カ国，31グループが登録し，28グループが結果を提出した．内訳は，随時検索17，言語横断検索10，用語抽出9グループ．検索系では，国内の企業研究所9，国内の大学・国立研究所10，海外4であり，計117通りの結果を提出した．国外研究機関との混成の国内グループもある．日本語を解するメンバなしで参加したグループが2つあった．

■ 日程： ■

1998年11月初旬　：文書データと訓練用課題とその正解文書リストを配布
1999年2月8日　：評価用検索課題の配布
3月5日　：検索結果提出
8月30～31日：成果報告会，会議録発行

参加グループは，文書データと訓練用の検索課題を用いて各自の検索システムを訓練し，2月に配布された新しい評価用検索課題を用いて検索を行い，その結果を事務局に提出した．それを集めて人間の判定者が正解・不正解の判定を行うことによって正解文書リストを作成し，提出された各チームの検索結果を評価した．対話型システムも参加可能とした．

■ テストコレクション1（NTCIR-1） ■

第1回NTCIRワークショップ使用した**テストコレクション1（NTCIR-1）**は文書，検索課題，正解文書リスト，タグ付きコーパスからなる．

(1) **検索対象文書（Document）**（図-4参照）：日本国内

の65学会の協力を得て，全国大会，研究会などの発表論文の要旨を集めた**学会発表データベース**から選択した約33万件の文書から特定項目を抽出した．約半数の文書は日英対訳である．1レコードは，標題，著者名，会議録名，学会名，発表年月日，要旨，著者キーワードを含む．

(2) **検索課題** (Search topics)（**図-5**参照）：検索課題は，利用者の検索要求を一定の書式で明文化したものである．訓練用30課題，評価用53課題がある．検索課題は，当該分野の研究者などの利用者から収集した．**現実の，自然な**検索課題の収集を意図し，検索システムに投入する**問合せ**ではなく，利用者の検索要求の自然言語による記述とした．

<検索要求>は，利用者の検索要求を記述した自然言語の文，<検索要求説明>は，背景説明，検索の目的，正解判定基準，用語の定義などを含み，検索要求の背景情報を提供する．対話型システムの評価では，擬似的な利用者プロファイルとなる．<タイトル>は，検索課題を簡潔に表現したもので，サーチエンジンなどにしばしば投入される非常に短い問合せに相当するものとして，検索実験に用いることもできる．検索実験では，検索課題のどの項目を使用してもよいが，使用した項目を報告する必要がある．

検索課題の選定条件は，正解文書が5件以上あり，簡単すぎない．すなわち，単純な検索語のマッチングだけでは全正解文書を検索できず，検索語を含む文書であっても不正解があり得るよう考慮した．現実の文書では，同一概念が多様な用語で表現され，同一の語が多様な概念を表すからである．

(3) **正解判定（正解文書の網羅的リスト）**：正解判定は2名のクロスチェックに基づき最終判定をした．判定は，検索要求に「適合」，「部分的適合」，「不適合」の3段階である．適合文書には，適合と判断する根拠となった語句や文を抽出して付記した．

(4) **タグ付きコーパス**：文書データの一部に，語構成要素まで考慮した詳細な形態素タグを付与した．用語抽出タスクで使用し，正解データとして専門用語と用語役割の候補リストがある．

■ テストコレクションの評価 ■

NTCIRでは，テストコレクションの評価用ツールとしての**信頼性，妥当性**について，以下の観点から分析を行った．

(1) 判定者間の正解判定の一貫性
(2) 正解判定の網羅性
(3) 検索課題ごとの分析（難易度などのバランス）

正解判定は，人間の行う主観的な作業であり，あいまい性が伴う．プーリングの網羅性も問題となる．テストコレクションは，同じコレクションを使用したシステム間の検索有効性の相対的な比較に用いるものであるので，正解判定の一貫性や網羅性が，システム間の比較に影響を及ぼさず，どのシステムにも公平であることが重要である．

そこで，各検索課題について，2名の判定者の判定と最終判定との3種類の正解判定について，システムの評価への影響を調べた．プーリングについては，各提出結果からプールする文書数を変化させて，システム評価への影響を調べた．その結果，プーリング法の違い，正解文書候補収集の網羅性の差違，複数の判定者間の同一検索課題についての正解判定の不一致にもかかわらず，どの方法で作成された正解文書リストを用いても，システム間の相対的な順位付けには，非常に高い相関がみられ，一貫した評価結果が得られることが示された．また，プーリングに参加しなかったグループについても，特に不利になるという結果は認められなかった．

正解判定の網羅性の問題は，システム間の相対的評価に影響がないというのが，そのテストコレクションが受容され得る1つの規準となるが，難しい問題であり，引き続き，検討が必要である．

■ 評価結果：随時検索： ■

結果の一部を以下に紹介する．詳しい結果については，会議録などを参照していただきたい[4]．

(1) 検索課題の多くの項目を使用した（問合せが長い）検索ほど，一般によい結果が得られた．
(2) 対話型検索は，非対話型システムよりも検索有効性が高かった．対話の効果と，人間の探索者がやりとりにかかわった程度はさまざまであった．
(3) 検索課題の<概念>を使用した検索は，使用しない場合よりも検索有効性が高かった．
(4) 検索式拡張は，複数のグループで用いられ，対話型・非対話型のいずれでも有効であった．

図-6は，随時検索タスクの上位の結果，**図-7**は，随時検索の中で，非対話型の短問合せ検索の上位の結果である．**短問合せ**は，検索課題の<検索要求>のみを用いた検索，**長問合せ**は，<検索要求説明><概念>なども用いた検索を指す．

Aは，高度な自然言語処理を用いたフレーズ型索引を用いたベクトル空間型モデルのシステムで，適合フィードバックは対話型・非対話型の双方で有効であった．検索結果の下位まで有効性が高い点が特徴的である．A3は対話型長問合せ，A2は非対話型の長問合せ，A1は非対話側の短

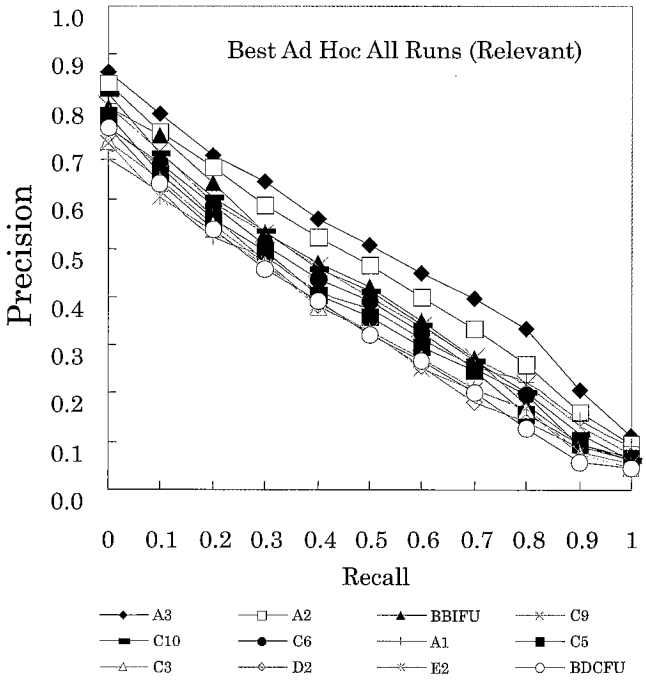
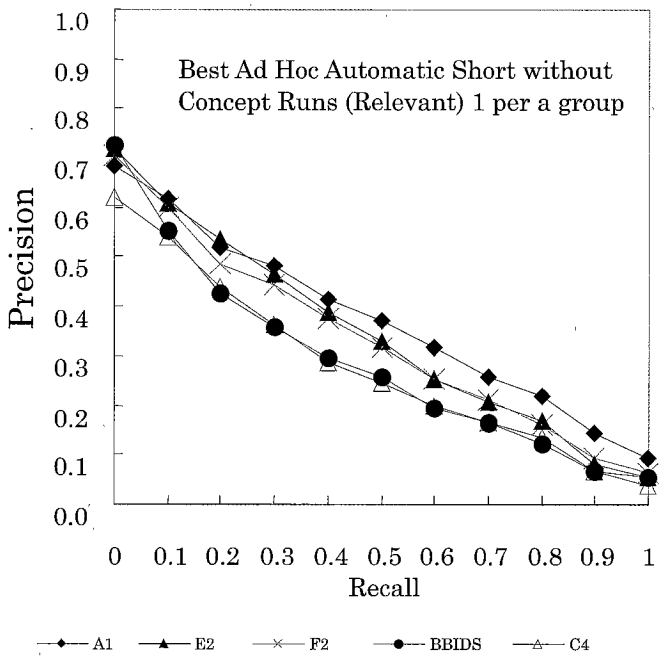


図-6　随時検索タスクの上位の結果



<検索要求>部分のみを使用した非対話型検索システム．1システムから1結果のみ．

図-7　随時検索タスクの上位の結果（2）

問合せである．

Bは，ロジスティック回帰を取り入れた独自の確率型モデルを用いた検索システムで，特に，上位の検索有効性が非常に高い点が特徴的である．BBIFUは，ひらがなを除外した非対話型のbi-gram索引の長問合せ，BDCFUは，辞書の最大一致法に基づく非対話型の単語，フレーズ型索引の長問合せ検索である．

Cは，Robertsonのモデルに基づく確率型モデルで，索引語はEDR辞書に基づく語，フレーズ分割である．C*は，いずれも非対話型の長問合せ検索．

Dは，ベクトル空間型で，重み付けにtf/idf，文書長のほか，共出現などを採用．語の分割は，辞書との極大一致法による．D2は，検索課題の<タイトル>と<検索要求>のみを使用した短問合せ．

Eは，ベクトル空間型の拡張として，Super imposeという問合せベクトルを重ね合わせて特徴を強化するモデル使用している．E2は，非対話型短問合せ．

Fは，B-tree上で拡張n-gramを用いていたベクトル空間型モデルのシステムであり，重み付けも，tf/idfに文書長を加えている．上位で有効性が高い．

興味深い点は，自然言語処理を活用した語・フレーズ分割を採用したシステムAと，ひらがなを除外した単純なbi-gram索引に確率型を拡張した検索モデルに重点を置いたシステムBとの双方がよい成績を収めた点である．検索モデルと重み付けについては，英文の検索で用いられていた方式を，日本語検索用にチューニングせずに用いたグループが少なくなかった．日本語文書の検索では，語分割が重要な課題の1つであったが，今後は，それとともに，検索モデルや重み付け方式自体にも，さらに関心が集まり，研究が進むことが期待される．

成果報告会

IREXとNTCIRの成果報告会は，連続開催した．

1999年8月30～31日	NTCIR	非公開　英語
1999年9月1日	NTCIR/IREX合同	公開　英語
1999年9月2～3日	IREX	公開　日本語

成果報告会の3日目は，NTCIRとIREXの合同ワークショップとし，基調講演に，米国の情報検索と情報抽出の評価ワークショップTRECとMUCの議長，Donna Harman女史（米国国立標準技術院：NIST）とRalph Grishman教授（ニューヨーク大学）を迎えた．パネルディスカッションでは，ACM-SIGIR2000実行委員長MunKew Leong博士（シンガポール国立ケントリッジデジタル研究所）の基調講演に続き，今後の評価ワークショップについて，活発な議論が交わされ，盛会のうちに散会した．

このような評価ワークショップは，しばしば，コンテスト，コンペティションといわれるが，これらの成果報告会では，興味深い多様な研究成果が発表され，順位付けへのこだわりはなく，参加者が互いに学び合う充実した会となった．

成果報告会については，公開形式，言語，提出結果の匿名性などについて，今後，さらに議論が必要であろう．

NTCIRの成果報告会に続けて開催された検討会では，Donna Harman，Fred Gey，MunKew Leong，岸田和明の各氏を中心に，今回のNTCIRワークショップの利点，欠点を議論した．そこでは，発表論文の質が高く評価され，また，対象言語や文書種類，検索と用語抽出の評価を一緒に行うことなど，ユニークな点が評価され，今後へ期待が寄せられた．また，国際会議の招待講演などにお招きいただく機会も少なくなく，日本語文書を対象とした評価プロジェクトが国際的な関心を集めていることがうかがわれる．

今後の展望

次回は，IREXとNTCIRが合同して，各サブタスクの企画の自主性を尊重しながら，NTCIRワークショップ2000/2001を開催する．NTCIRグループは主に情報検索，IREXグループは自動要約のタスクを企画運営する．詳しくはhttp://www.rd.nacsis.ac.jp/~ntcadm/workshop/を参照されたい．

2000年5月　：参加者募集
6月　：訓練用データ配布
8月上旬　：評価用データ／検索課題配布
9月中旬　：評価用結果提出
2001年2月下旬　：成果報告会

情報検索コミュニティは，その時々の要請に応じた研究開発に必要な実験用資源の整備と評価手法の検討を進めてきた．第1回IREXおよびNTCIRは，標準的な評価手法に準拠することに重点をおいてきた．今後は，以下の点に考慮し，より有効な評価のあり方を検討していきたい．

1. 文書種類，規模の拡充
2. より自然な検索課題
3. 対話型システムへの対応
4. 多言語，言語横断への対応
5. 文書の検索から，テキスト利用支援へ
6. 評価法・評価尺度の検討

現実に即した文書や検索課題の収集については，著作権など微妙な問題もある．実用検索サービスやその他多くの電子文書を生産，管理している部門からの幅広い理解と協力を期待したい．それとともに，これらの研究資源を使用する研究者一人一人が著作権に十分配慮をして研究を行い，著作権者からの信頼を積み重ねていくことも重要である．

対話型については，利用者側の問題，検索された文書の拾い読み，自動要約，回答抽出，複数文書の対比などの検索結果文書の利用を助ける技術など多くの興味深い課題がある．

多言語対応では，より自然な検索課題と正解判定について，国際的協力が必要である．NTCIRは，TREC，韓国，台湾，欧州のCLEFなどのプロジェクトとも密接な連係，協力関係を保っている．

現在の情報検索システムは，利用者にとって必要な情報を含む可能性が高い文書を検索する．それだけでなく，文書から必要な情報そのものや回答の抽出，検索された文書の拾い読み，要約，分析，対比などの情報活用を支援する機能も必要である．このような検索後の処理や質問回答などには自然言語処理技術が重要であり，情報検索と自然言語処理技術の融合が新たな情報検索・テキスト情報利用支援システムの可能性を開く．

IREXと合同した拡大NTCIRワークショップは，今後，伝統的な情報検索システムの検索有効性の評価法を踏まえつつ，自然言語処理との関連も含め，新しい課題も積極的に取り上げ，その「評価」に関する検討をしていくことが情報検索・テキスト利用技術の一層の発展につながると期待している．

■ NTCIRワークショップへのお問合せ ■

〒101-8430 東京都千代田区一ツ橋2-1-2
学術総合センタービル内
学術情報センター研究開発部
神門典子（かんどのりこ）　Email: ntcadm@rd.nacsis.ac.jp
電話: 03-4212-2529　FAX: 03-3556-1916

謝辞　NTCIRは，学術振興会未来開拓研究事業JSPS-RFTF96P00602の研究資金によります．NTCIR，IREXとも，情報処理学会協賛，情報学基礎研究会後援をいただき，関連して，1999年前期全国大会の公開パネル，情報学シンポジウムなど発表の機会をいただきました．感謝申し上げます．

参考文献

1) TREC: http://trec.nist.gov/
2) Sakai, T. et al.: BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems, SIGIR-Forum, Vol.33, No.1, pp.13-17 (1999).
3) IREX: http://cs.nyu.edu/cs/projects/proteus/irex/
4) NTCIR: http://www.rd.nacsis.ac.jp/~ntcadm/
5) NTCIR 2000/2001: http://www.rd.nacsis.ac.jp/~ntcadm/workshop/
6) 神門典子: 情報検索システムの評価を巡って: テストコレクションとコンペティションを中心に, 1999年情報学シンポジウム講演論文集, pp.129-136 (1999).
7) Spark-Jones, K. et al.: Report on the Need for and Provision of an 'ideal' Information Retrieval Test Collection, Computer Laboratory, Univ. Cambridge (1975).
8) TDT: http://www.itl.nist.gov/iaui/894.01/tdt3/
9) CLEF: http://www.iei.pi.cnr.it/DELOS/CLEF/

(平成12年2月23日受付)