

## 情報アクセス技術の性能評価のための研究基盤 NTCIR : 言語横断検索の創成と展開

神門 典子<sup>\*</sup>, 岸田和明<sup>†</sup>, 栗山和子<sup>‡</sup>, KH Chen<sup>§</sup>, HH Chen<sup>§</sup>,  
Sukhoon Lee<sup>\*\*</sup>, Sung Hyon Myaeng<sup>††</sup>

本稿では、情報検索・要約・情報抽出・質問応答・テキストマイニングなどの情報アクセス技術の評価を行なう NTCIR(エンティサイル)プロジェクトと NTCIR ワークショップの概要を紹介する。評価はどのような技術の研究開発においても必要不可欠なものであるが、情報アクセス技術では、成否の判定基準として利用者の判断が用いられ、多くの不確定な要素を含み、評価の実施には多くの課題がある。NTCIR では 1 年半に 1 回のサイクルでワークショップを企画し、国内外の多数の研究グループが集まって共通の研究課題を設定し、大規模なテストコレクション（評価実験用の正解データ付きデータセット）を構築し、評価手法を提案し、その共通基盤の上で多数の研究グループが比較評価し、互いに学びあいながら研究を進めてきた。本稿では、その概要、背景、経緯、構築したテストコレクションなどを概観するとともに、言語横断検索タスクについて概観し、今後の方向性を展望する

### NTCIR: Research Infrastructure for Evaluation of Information Access Technologies — Focus on Cross-Lingual Information Retrieval --

Noriko Kando<sup>\*</sup>, Kazuaki Kishida<sup>†</sup>, Kazuko Kuriyama<sup>‡</sup>, Kuang-hua Chen<sup>§</sup>, Hsin-Hsi Chen<sup>§</sup>, Sukhoon Lee<sup>\*\*</sup>, Sung Hyon Myaeng<sup>††</sup>

This paper provides an overview of the NTCIR project, which has constructed research infrastructure of comparative evaluation of information access technologies, such as information retrieval, summarization, information extraction, question answering, text mining, etc.. The main components of the infrastructure are: large-scale reusable test collections usable for experiments, evaluation methodologies and metrics and community of the researchers who are interested in exchanging research idea in informal atmosphere and research boosting via comparative evaluation of the technologies. It outlines a brief history of NTCIR, test collections constructed and focusing on the Cross-Lingual Information Retrieval. Finally some thoughts on the future directions are provided.

---

<sup>\*</sup> 国立情報学研究所, National Institute of Informatics

<sup>†</sup> 慶應義塾大学, Keio University

<sup>‡</sup> 白百合女子大学, Shirayuri University

<sup>§</sup> 国立台湾大学, National Taiwan University

<sup>\*\*</sup> National Chungnam University

<sup>††</sup> International Communication University

## 1.はじめに

情報検索・要約・情報抽出・質問応答・テキストマイニングなどの情報アクセス技術の研究では、大規模なデータセットを用いた実験によって新しく提案した手法が既存のものより優れていることを示す必要がある。しかしながら、情報アクセス技術の評価にはさまざまな不確定な要素が含まれ、信頼性、感受性、頑強性の高い評価をしていくか、多くの課題がある。

われわれは NTCIR(エンティサイル)<sup>7</sup> [1] プロジェクトを通じて、検索・要約・情報抽出・質問応答・テキストマイニングなどのさまざまな情報アクセス技術を取りあげ、その有効性を評価する基盤を構築している。その研究基盤は、「テストコレクション」とよぶ評価実験用の正解データ付きデータセット、評価手法や評価指標などであり、それをささげるのは共通基盤の上で研究を進める研究者コミュニティである。

プロジェクトは 1997 年末に開始した。以降、概ね 1 年半に 1 回のサイクルで、国際ワークショップを企画し、各回、複数の技術分野を選び、国内外の多数の参加研究グループが共通のテストコレクションを用いて実験を行い、実施結果を集めて正解データを作成し、システムの比較評価、研究アイデアの交換などにより、その技術分野の研究を Leverage することをめざしている。<sup>8</sup>

### 1.1 情報アクセス技術と評価

情報検索は、インターネットサーチエンジンとして多くの方々にとって身近な社会基盤の一つとなっている。これは蓄積された大量の文書から利用者の情報要求にレバントな(適合する)情報を選び出す技術である。伝統的にはレバントな情報を含む文書を探し出す「文書検索」である。しかし、利用者は検索したあと、検索された文書を読んで適合性を判定し、必要な情報が得られる箇所を探しだし、読み、理解し、比較や分析をして、新たな意思決定をしたり、レポートを書くなどのなんからのタスクを行う。また、技術の進歩と社会の発展のなかで、次第に検索技術に求められるものも多様化してきている。

NTCIR では、プロジェクト開始当初から、図 1 にしめすように、伝統的な情報検索の評価実験とともに、新たなチャレンジという 2 つの柱をた

てて研究を進めてきた。

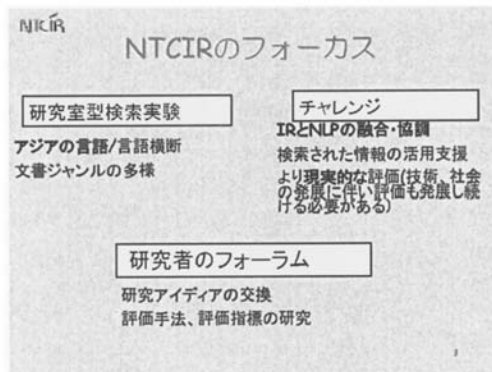


図 1 NTCIR のフォーカス

情報検索では、とくに、自国語である日本語とその周囲にある東アジア言語、言語の壁を越えて情報にアクセスする言語横断情報検索、そして、多様な文書ジャンルを対象とした検索技術の研究に焦点をあててきた。

他方、チャレンジとしては、自然言語処理と情報検索技術の連携、融合による新たな技術分野の展開を目指していた。具体的には、要約、質問応答、テキストマイニング、情報抽出、可視化など、検索された文書中の情報の活用を支援する情報アクセス技術を研究対象としてきた。さらに、より現実的な、信頼性、感受性、頑強性の高い評価実験手法や指標の研究、評価における振る舞いなどについても研究を進めてきた。

それらの 2 つの活動を支えるのが研究者のフォーラムである。

さらに、今後は、チャレンジの方向として、検索を基盤として、膨大な情報空間の中を利用者を導き、新たなつながりを見出す、情報から知を紡ぎ出す方向への展開を目指している。

### 1.2 NTCIR ワークショップ

NTCIR では、いままでに図 2 のような研究部門(「タスク」という)を取り上げてきた。図 3 に示すように新しい研究部門の導入により新たな研究者コミュニティへと参加者へとひろがってきた。

<sup>7</sup> NII Test Collection for Information Retrieval Systems.

詳しくは <http://research.nii.ac.jp/ntcir/> をご覧ください

<sup>8</sup> 研究対象は、主に、東アジア言語(日本語、中国語、韓国語)であるが、欧米からも多くの参加があり、毎回、約 15 カ国から 100 以上の研究グループが参加している。

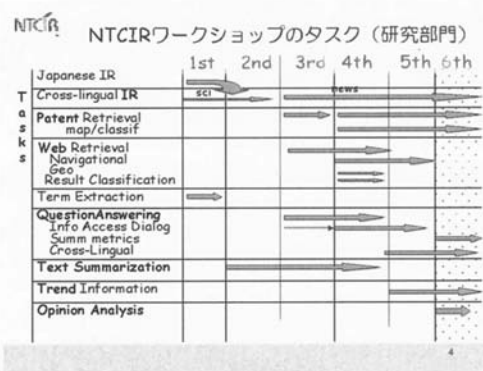


図2 NTCIRワークショップのタスク

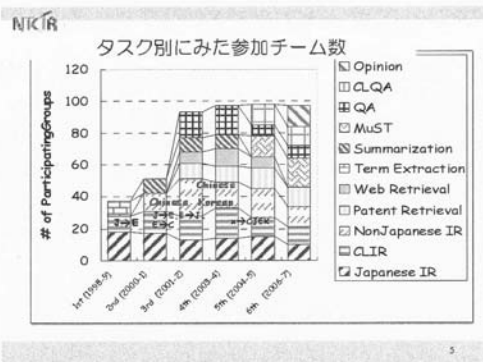


図3 タスク別にみた参加チーム数

検索では、初期には、学術文書の検索と言語横断検索から始まり、新聞記事の言語横断検索、さらに、特許やWEB、Blogなどの特定の文書ジャンルの検索へと発展してきた。文書ジャンルの拡大は、その背後にある想定される利用者コミュニティ、そこでの典型的な使用の状況や情報探索タスクが異なり、それぞれの状況に適した実験設計と判定基準を設定することで、研究室内での実験であっても、より現実に意味のある評価実験の遂行を心がけた。また、文書ジャンルの拡大は、用語や文書長などの多様性にもつながった。

このうち、特許検索・特許マップ自動生成などのテキストマイニング・特許の自動分類については藤井ら[2]が、質問応答・言語横断質問応答については森ら[3]が、意見分析については関ら[4]が報告する。言語横断検索については後述する。

以下、次節では、NTCIRで構築したテストコレクション、3節では、情報アクセスの評価をめぐる問題を概観する。4節では、いままでのNTCIRで行ってきた、言語横断検索の評価に業焦点をあてて、研究活動を紹介する。最後に簡単に今後の展望を述べる。

## 2. NTCIR テストコレクション (付表参照)

これまで、一連のNTCIRワークショップを通じて、付表に示したテストコレクションを構築してきた。その多くは、ワークショップ後、参加者以外の研究者にも研究目的利用で公開している。

テストコレクションとは、評価実験に使用する文書データセット、質問や課題のセット、各質問や課題に対する正解データの3つからなる。正解データは人手で判定するため、多くのコストがかかるが、一旦構築すれば繰り返し利用することができる。情報検索テストコレクションは、欧米では1960年代から構築と共有が進められてきた長い歴史がある。わが国や東アジアでは1990年代半ばに情報処理学会のWGで構築しBMIRが、多くの研究者が共通に使えるテストコレクションの最初であろう。その後、日本では、検索と情報抽出を対象としたIREXとNTCIRが同時に開始され、NTCIR-2からは、両者のメンバが合同し、協力してNTCIRワークショップを企画運営している。

テストコレクションを用いた実験は、実際の利用者が行う情報アクセスを有効性の評価に関わる最も中心的な要素だけを取り出して実験するもので、いわば、現実を抽象化したものと考えられる。

正解判定は人間が行う。そのため、主観的で、揺れがあり、手間がかかる。判定に揺れがあっても安定する評価法が必要である。また、文書データセットや課題によって結果が異なるので、同一テストコレクションを用いたシステム間の相対比較によって評価する。複数のテストコレクションを用いて事件をすることでより信頼性の高い評価結果を得ることができる。

また、テストコレクションを用いた評価実験は唯一の評価方法ではない。いわば、薬の開発過程になぞらえたら、”in vitro(試験管内での)”実験であって、研究開発の初期にたくさんの研究アイデアの仲から有望なものをテストする場であり、情報アクセス技術の研究開発では必要不可欠なものである。

NTCIR テストコレクションの研究目的利用数は1002件(2007年3月末現在、現在、有効な覚書数による)である。

## 3. 情報アクセスの評価

情報アクセス技術の評価には、さまざまな観点がある。たとえば、Cleverson & Keenは図4のような6つのレベルを提唱している。テストコレクションを用いた評価実験では、主に、この中の「処理レベル」、すなわち、提案するアルゴリズムや

メカニズムの有効性、すなわちどのくらいうまく情報にアクセスができたか、どのくらい無駄なく、もれなく必要な情報を求める形で得ることができるかを評価する。

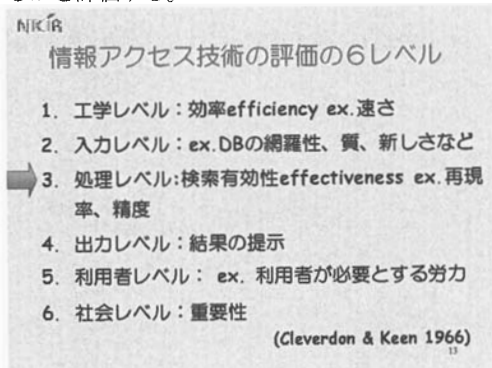


図4 情報アクセス技術評価の6レベル

コンピュータを使用した情報検索や要約などの情報アクセス技術の研究は1950年代に開始され、「人間の主観的判定」を成否の基準にした最初の計算機科学の分野といわれている。また、現実に情報アクセスの対象となる文書データが増大、多様化するなかで、評価実験も規模と多様性が求められる。

また、情報アクセス技術は、一般に、使用するデータセットや質問によって難しさが異なるため、評価指標の絶対値には意味がなく、同じデータセットを使用したシステム間の相対比較によって評価を行う。

図5に示すように、システム間の差よりも同一システム上での検索質問間の差の方が大きい。しかも、どの質問が難しいかは、システムごとに異なる。より信頼性の高い評価を行うには、大規模なデータセットとより多数の多様な質問を用いる必要がある。また、使用する評価指標によっても、どれだけの数の質問が必要であるかは異なってくる。人間による判定はコストが高く、個別の研究グループでは十分な規模の実験用データセット作成は困難である。

他方、一旦、作成した正解付きの実験用データセットと評価指標といった評価実験の基盤があれば、研究アイデアをすぐに実験で試すことで、研究アイデア交換を促進し、ビジビリティや関心を高め、領域への新規参入者にもわかりやすく、すその拡大にも貢献するなど多くの意義が考えられる。

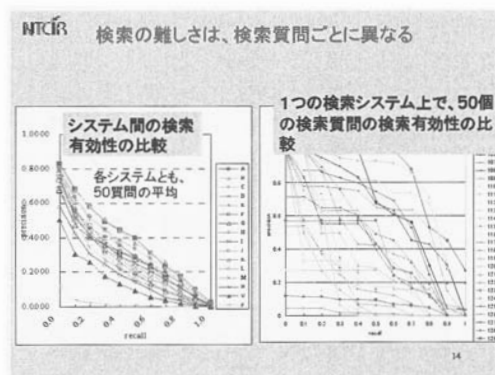


図5a システム間の差よりも同一システム上での検索質問間の差の方が大きい。

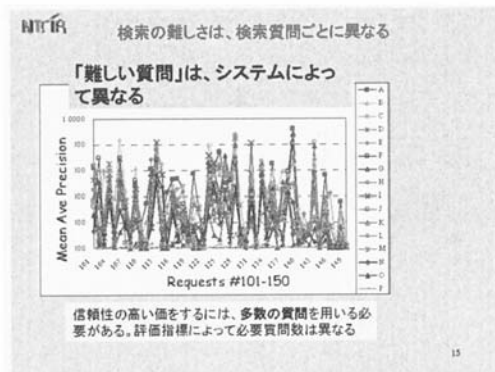


図5b 「難しい質問」はシステムによって異なる

テストコレクションはなぜ必要か？ 情報アクセスの有効性は使用するデータセットによって異なり、共通の基盤が必要である。また、再現率；現実の場では評価できない。利用者を含む評価は高価で、かつ、統制できない要因が多く、結果の解釈が難しい。技術の核となる機能について、わかりやすい評価が必要である。特に研究開発の初期には実験室で繰り返し研究アイデアの有効性を評価し、有望な要素技術や改善を積み上げていくことが必要である。テストコレクションはこのような実験に新たに被験者を招請することなく、簡単に繰り返し使用することができると考えられている。



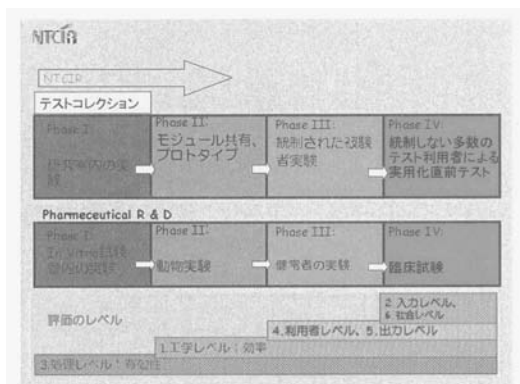


図6 検索フェーズと評価

他方、テストコレクションを用いた実験は、現実の情報アクセスから、さまざまなコントロールしにくい要素を捨象して、技術のもっともカーネルな部分の有効性を、繰り返し、実験できるようにするためのものである。

情報アクセス技術の研究開発の過程を、よく、製薬企業での研究開発になぞられる。製薬の研究開発では、まず、フェーズ1では、多数の物質について試験管内での実験を多数、重ね、非常に有望な物質のみが、より高価で生命の危険にも関わる、フェーズ2の動物実験に進むことができる。さらに、そこで有効性、安全性が確かめられるとフェーズ3の健常者での実験、最後に患者さんを対象とした臨床試験をおこない、実用化へと進む。

同様に、情報アクセス技術の研究開発でも、最終的には、実用化を目指すものであり、現実的な環境で利用者を多数用いた評価も必要である。しかし、研究の初期には、多数の、さまざまな研究アイデアをすぐに実装して、その有効性を評価でき、繰り返し、いろいろなアイデアをためせることが重要である。テストコレクションは、そのようなフェーズで、とくに、研究室に多数の被験者を集めなくても繰り返し実験できることがとくに意義のあることである。これによって、研究を効率化、高速化、さらにコミュニティ創成などの効果も期待される。

また、多数のテストコレクションの構築、評価ワークショップ、評価手法についての研究が進む中で、次第に、情報アクセスシステムの振る舞い、評価指標の振る舞いや特徴、利用者実験との関係など多くのことが明らかになってきて、さらに、分野の推進に貢献している

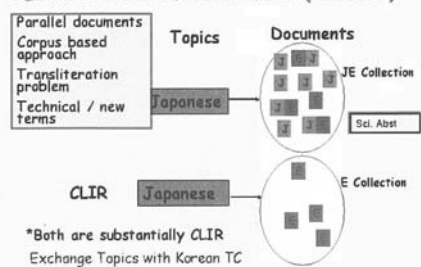
#### 4. 言語横断検索

NTCIR ワークショップの歴史の中で、言語の

壁を越えて、他の言語で書かれた情報にも一括してアクセスする言語情報検索、言語情報アクセスは、NTCIR-1 から、継続して、研究を進めてきた部門である。

図7に示したように、NTCIR-1と2では、学術文書を対象とした、日英の言語横断検索をとりあげた。ここでは、まず、日本語の索引方式について多くの議論が戦わされた。また、学術用語は日本語訳、英語等の原綴、そのカタカナ読みなど多様な表記があり、そのような表記のゆれへの対処法が課題の一つとなった。また、学術用語等の翻訳なども研究された。

CLIR at NTCIR: NTCIR WS 1 (1998/99)



NTCIR WS 2: Japanese & English

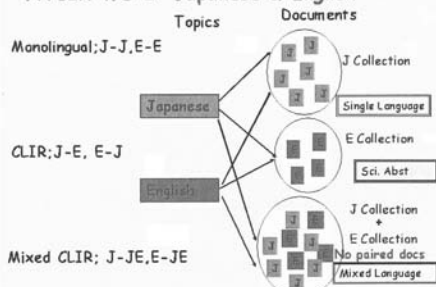
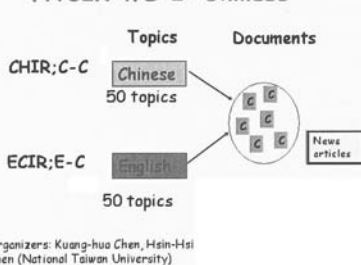


図7 学術文書の言語横断検索

NTCIR WS 2: Chinese



Organizers: Kuang-huo Chen, Hsin-Hsi Chen (National Taiwan University)

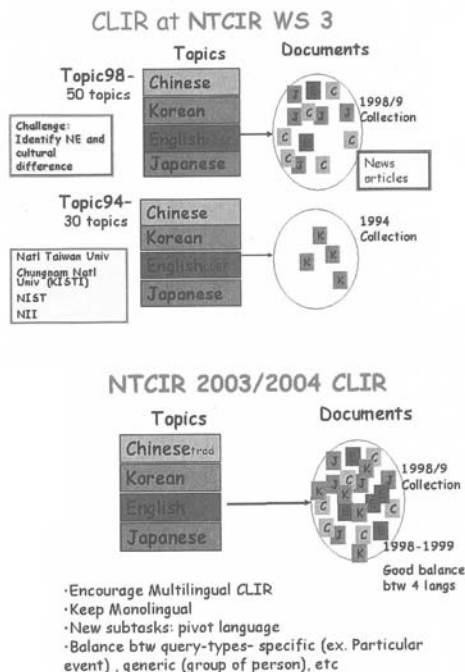


図8 新聞記事を対象とした言語横断検索

NTCIR-3以降は、新聞記事を対象とした、日中韓英の4カ国語の言語横断検索をとりあげた。検索課題は、各言語からあつめて共通のものを使用した。ここでは、とくに固有名、文化的・社会的背景が大きな課題となり、翻訳辞書にのっていない多数のOut-of-Vocabulary(OOV)をどのように解消するか、さまざまな手法が提案された。たとえばふだんに翻訳資源がある言語を中間言語として使用するPivot検索、WEBの使用などである。欧米など同一アルファベットを使用する言語間では、固有名詞はもつとも簡単なものであるが、東アジアのように、それぞれ異なる文字体系を持つ言語間では、固有名詞がもつとも困難な課題の一つであり、その解消に向けて多くの研究が進んだ。

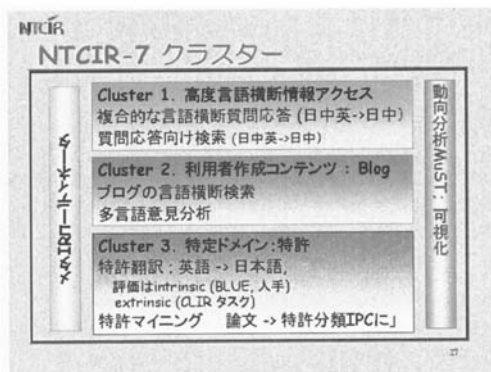
また、特許と新聞のジャンル横断、特許の言語横断検索なども、文書長の課題、高度に専門的な用語、構造化した文書の検索など新たな課題を提供した。さらにNTCIR-5からは、固有名を中心とした言語横断質問応答に取り組んだ。

## 5. 今後の展開

現在、参加者募集中のNTCIR-7からは、タス

クの構成を大きく変え、基盤技術としての検索とそこから情報を引き出し活用するための諸技術を組み合わせた研究と評価を可能なようにしている。(図9参照)。より多くの方の参加と議論を期待している。

NTCIRの次の10年間の目標は、情報から知を紡ぎだすための技術である。個別の情報の検索だけではなく、情報空間のなかでのつながり、行動の履歴、関係付け、利用者を含む探索の過程とその支援技術、簡単な辞書引きのようなlook upではなく、より大きな問題の解決に資する情報アクセス技術。また、ノービスな利用者を支援する技術などを対象としていきたい。技術は常に進歩する。社会も常に進歩する。評価は、そのような技術と社会の進歩に即して発展していくべきものである。



- [1] Kando, N. Evans, D.E. (eds). NTCIR-6: Proceedings of the 6<sup>th</sup> NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access. May 2007, NII, Tokyo.
- [2] 藤井敦ほか. 特許情報処理を指向したテストコレクションの構築: 情報検索と自然言語処理の融合を目指して. 情報処理学研究報告. 2007-FI-89, NL-183 (to appear in this volume)
- [3] 森辰則ほか. NTCIRにおける質問応答技術の評価と今後の展望. 情報処理学研究報告. 2007-FI-89, NL-183 (to appear in this volume)
- [4] 関洋平ほか. 意見分析タスク-多言語テキストを対象とした意見抽出技術の評価. 情報処理学研究報告. 2007-FI-89, NL-183 (to appear in this volume)

NTCIRテストコレクション詳細一覧

NTCIR テストコレクション : IR & QA

分類	コレクション	タスク	文書データ						タスクデータ		
			ジャンル	ファイル名	言語	年産	文書数	サイズ	検索課題/質問 言語	#	適合判定
OLIR on Scientific	NTCIR-1	IR	科学技術 文献抄録	ntc1-jc(A)	JE		338,483	577MB	J	83	3 grades
				ntc1-j(A)	J	1986-	332,518	312MR			
				ntc1-c(A)	L	1987	187,060	218MB		60	
				ntc1-tmrc(A)	J		2,000				
	NTCIR-2	IR	科学技術 文献抄録	ntc2-j(A)	J	1986	400,248	600MB	JE	49	4 grades
ntc2-c(A)				E	1989- 92	134,578	200MB				
OLIR on News	CIRB010	IR	新聞	CIRB010(C)	Ct	1988- 1989	132,220	132MB	CtE	50	4 grades
	NTCIR-3 CLIR	IR	新聞記事	KLBB010(C)	K	1994	66,146	74MB	CtKJL	30	4 grades
				CIRB011(C)	Ct		132,173				
				CIRB020(A)			249,508				
				Mainichi(H)	J	1988- 1989	270,078	870MB			
				EBR010(C)			10,204				
	NTCIR-4 CLIR	IR	新聞記事	Mainichi Daily (A)	E		12,723		CtKJE	60	4 grades
				CIRB011(C)	Ct		132,173				
				CIRB020(A)			249,203				
				Hankook(A)	K		149,521				
				Chosenilbo(A)			104,517				
				Mainichi(B)	J		223,078				
				Yomiuri(B)			373,558	ca.3GB			
				EBR010(C)		1998 1999	10,704				
				Mainichi Daily (A)			12,723				
				Korea Times(A)			19,959				
	NTCIR-6 CLIR	IR	新聞記事	Hong Kong Standard(A)			96,863		CtKJE	50	4 grades
Xinhua(B)						208,167					
CIRB040(A)				Ct		801,446	581.7MB				
Hankook(A)				K		85,250	52.1MB				
Chosenilbo(A)						135,124	88.7MB				
Mainichi(H)				J		199,681	118.8MB				
Yomiuri(B)						658,719	343.3MB				
NTCIR-8 CLIR	IR	新聞記事	Mainichi Daily (A)		2000- 2001	12,155	9.9MB	CtKJE	50	4 grades	
			Korea Times(A)			30,530	25.3MB				
			Daily Yomiuri (B)			17,741	22.9MB				
			Xinhua(B)			198,674					
			CIRB040(A)	Ct		801,446	581.7MB				
NTCIR-9 CLIR	IR	新聞記事	Hankook(A)	K	2000 2001	85,250	52.1MB	CtKJE	50(selected from NTCIR-3,4)	4 grades	
			Chosenilbo(A)			135,124	88.7MB				
			Mainichi(H)	J		199,681	118.8MB				
			Yomiuri(B)			658,719	343.3MB				
			CIRB040(A)	Ct		801,446	581.7MB				
CLQA	NTCIR-5 CLQA	QA	新聞記事	Yomiuri(H)	J	2000- 2001	658,719	343.3MR	CtJ	smpl:300, test:200*6	3 grades*7
				Daily Yomiuri (B)	E		17,741	22.9MB			
	NTCIR-8 CLQA	QA	新聞記事	CIRB020(A)	Ct		249,203	320MR	CtE	J-E/J-J/E-J-200 C-E/C-C/E- C/L-L:100	3 grades*7
				Mainichi(H)	J		220,078	282MB			
OPINION	NTCIR-4 OPINION	II / analysis	新聞記事	EBR010(C)			10,204	24.5MB	CtJ	3?(selected from NTCIR-3, 4-5 CLIR)	480*8  168*8  2 types, 3 metrics
				Mainichi Daily (A)	E	1988 1989	12,723	33.3MR			
				Korea Times(A)			19,959	55.8MB			
				Hong Kong Standard(A)			96,863	252MB			
				CIRB020(A)	Ct	1988- 1989	249,508	788MB			
				CIRB040(A)		2000 2001	801,446				
				Mainichi(B)	J	1988- 2001	419,759	766MR			
				Yomiuri(H)		1988 2001	1,004,699				
				Daily Yomiuri (B)		2000- 2001	17,741				
				Mainichi Daily (A)		1988 2001	24,878	471.6MR			
Korea Times(A)		2000- 2001	30,530								
			Hong Kong		1988						

																		Standard(A)					
																		Xinhua(B)					
Patent	NTCIR-3 PATENT	IR	特許全文	kkh(A)*3	J	1998-1999	697,262	18GB	CtCsKJE	31										3 grades			
			特許抄録	jsh(A)*3 pai(A)*3	E	1995-1999	1,706,154 1,701,339	1.883MB 2,711MB															
	NTCIR-4 PATENT	IR	特許全文	Publication of unexamined patent application(A)	J	1993-1997	ca.1,700,000	ca.45GB	E												Main:34, Add:68	3 grades	
			特許抄録	Patent Abstracts of Japan(PAJ)(A)	E	1993-1997	ca.1,700,000	ca.2.2GB															
	NTCIR-5 PATENT	IR	特許全文	Publication of unexamined patent application(A)	J	1993-2002	3,498,252	94.5GB	JE													34+1189 in NRCIR-5, added 349+1681 in NTCIR-6	3 grades
			特許抄録	Patent Abstracts of Japan(PAJ)(A)	E	1993-2002	3,498,252	ca.5GB															
	NTCIR-6 PATENT	IR	特許全文	Patent grant data published from USPTO(A)	E	1993-2000	981,948		E													3221	3 grades
	QA	NTCIR-3 QA	QA	新聞記事	Mainichi(B)	J	1998-1999	220,078	260MB	J*1		1200											exact answer
		NTCIR-4 QA	QA	新聞記事	Mainichi(B)	J	1998-1999	220,078	ca.776MB	J*1		197											exact answer
					Yomiuri(E)			373,558			251												
		NTCIR-5 QA	QA	新聞記事	Mainichi(B)	J	2000-2001	199,581	260MB	J*1		50 series(3500)											graded
	NTCIR-6 QA	QA	新聞記事	Mainichi(B)	J	1998-2001	419,759	635MB	J		100Q(any kind of Q)											graded(3 types, 4levels)	
WEB	NTCIR-3 WEB	IR	Web (html/text)	NW100G-01(A)	multiple*4	crawled in 2001	11,038,720	100GB	J*1		47											4 grades +relative	
	NTCIR-4 WEB	IR	Web (html/text)	NW100G-01(A)	multiple*4	crawled in 2001	11,038,720	100GB	J*1													3 grades	
	NTCIR-5 WEB	IR	Web (html/text)	NW100G-04(A)	multiple*4	crawled in 2004	98,870,352	1.36TB	J*1		269+847											3grades	
MuST	NTCIR-6 MuST	IE/analysis	新聞記事	Mainichi(B)	J	1998-1999	220,078	260MB	J		27										581*9		
Others	available for future task		QA site on Web	Yahoo! QA corpus (Chiebukuro)(A)	J	Apr.2004 to Oct.2005																	
			News	Singapore Press(A)	Cs	1998-2001																	

J:日本語 E:英語 C:中国語 (Ct:繁体字 Cs:简体字) K:韓国語

- \*1: 英訳あり
- \*2: gakkai サブファイル: 1997-1999 kaken サブファイル: 1986-1997
- \*3: kkh: 未審査特許申請 jsh: 日本特許抄録 pai: jsh英訳
- \*4: 大部分は日本語または英語(一部他言語あり)
- \*5: 用語抽出/役割分析
- \*6: 中国語文書に対して300\*200質問、日本語・英語文書に対して300\*200質問
- \*7: Right, Unsupported, Wrong
- \*8: 意見情報をタグ付けた文書数 (文の数は中国語: 11,907、日本語: 15,279、英語: 8,628)
- \*9: 意向情報をタグ付けた文書数

NTCIR テストコレクション: Summarization

コレクション	タスク	文書データ				要約データ			
		ジャンル	ファイル名	言語	年度	文書数	type	analyzeta	total#
NTCIR-2 SUMM	単一文書 新聞記事 複数文書	新聞記事	Mainichi(B)	J	1994,1995,1998	180 doc	7	3	3780
NTCIR-2 TAD*10			Mainichi(B)		1998	1000 doc	2	1	2000
NTCIR-3 SUMM			Mainichi(B)		1998-1999	60 docs	7	3	1260
						50 sets	2	3	300

J:Japanese

\*10: 現在配布を中止しています。配布が可能になるようでしたら、NTCIR-より e-mail にてお知らせします。  
(お知らせを希望される方は、NTCIR mailing list (「NTCIR一般メーリングリスト」)にご登録ください。)

(A)	NIJから研究目的でデータ提供
(B)	NTCIRワークショップの当該タスク参加者にはNIJから提供 参加者以外は研究目的で他の機関から入手可能参加者のみデータ提供
(C)	NTCIRワークショップ参加者のみデータ提供