LD_007

NTCIR公式結果に基づく文書検索技術の進歩に関する一考察

A Note on Progress in Document Retrieval Technology based on the Official NTCIR Results

酒井 哲也†
Tetsuya Sakai

1. はじめに

国立情報学研究所は1999年より約一年半に一度のペースでアジア言語情報アクセスに関する評価型ワークショップNTCIR(エンティサイル)を開催している[3]. このプロジェクトは，世界各国の研究者に一定期間共通のタスクに取り組ませることにより技術開発を加速させ，同時に，研究者間で共有し再利用できるテストコレクションを構築するという意義をもつ. しかし，例えば最近開催された第5回NTCIRの(単言語検索を含む)言語横断検索タスクには15カ国から100チーム以上の参加があり，各チームは独自の問題意識に基づいて実験を行い結果を報告しているため，主催者・参加者のいずれにとってもNTCIRプロジェクト全体を俯瞰して知見を得ることが容易ではなくなって来ている. NTCIRに先駆け，主に英語文書を対象として米国で開催されてきた評価型ワークショップTREC(Text REtrieval Conference)では，参加チームの検索有効性の一年毎の推移の調査などの分析が小規模ながら行われてきたが[2, 13]，NTCIRに関してはまだ歴史が浅いこともあり，同様な試みが(少なくとも論文等の明示的な形では)これまで報告されていない.

本研究の目的は，NTCIRの結果から文書検索技術の進歩が読み取れるかどうかを検証することである. 特に日本語検索に着目し，対比のために中国語検索についても補助的な分析を行う. さらに，単言語検索に加えて言語横断検索の進歩についても考察する. 分析対象としては，NTCIR-3(2002年10月)およびNTCIR-5(2005年12月)の言語横断検索タスクに提出された検索結果を用いる.(目下，これらはNTCIRから公開されている唯一のデータである.) すなわち本研究は，最近約3年の間に検索技術の進歩があったか否かを問うものである.

2. 分析対象データ

表1に，分析対象としたNTCIR-3・NTCIR-5言語横断検索タスクの日本語・中国語文書テストコレクションおよび各タスクに提出されたDESCRIPTION(検索課題を簡潔な一文で表現したもの)に基づく検索結果に関する情報を示す[3]. DESCRIPTIONに基づく検索結果はNTCIR-3と5の両方で提出が必須とされていたため今回の分析対象とした. 複数の検索結果を提出したチームについては，最も成績のよいもののみを分析対象とした.

NTCIR-3と5の両方において日本語単言語検索の結果を提出し且つチームIDを変更していないチームは4チームあった. 中国語単言語検索についての同様なチームは3チームあった. また，同一の検索課題言語を用いた日本語文書対象の言語横断検索結果をNTCIR-3と5の両方に提出したチームは2チームあったが，中国語文書については共通のチームがなかった.

表1: 分析対象としたNTCIR-3・5言語横断検索タスクのテストコレクションおよび検索結果データ

	日本語文書		中国語文書	
	NTCIR-3	NTCIR-5	NTCIR-3	NTCIR-5
検索課題数	42	47	42	50
文書数	220,078	858,400	381,681	901,446
高適合文書数	330	149	882	350
適合文書数	1,324	1,963	1,046	1,535
部分適合文書数	884	2,078	1,356	1,167
単言語チーム数 (DESC. のみ)	11	23	8	22
言語横断チーム数 (DESC. のみ)	6	16	5	7

3. 単言語検索の進歩に関する考察

3.1 NTCIR-3・5の公式結果概観

図1にNTCIR-3と5の日本語単言語検索結果をRelaxed Average Precision (AveP)[3]に基づき順位づけした結果を示す. また，AvePは高適合・適合・部分適合文書を同一視してしまうため，適合レベルを考慮しかつ信頼性の高い評価尺度であるQ-measure[9, 11]により同様の順位づけを行った結果を図2に示す. 両図におけるチーム順位は同一であり，グラフの形も似ている.(ただし，チーム内順位は異なる場合がある. 例えば，チームTSBのNTCIR-3における検索結果IDは両図で異なっている.) 比較のために中国語単言語検索結果についての同様のグラフである図3と図4を見てみると，こちらではNTCIR-5のチーム順位が異なっている. 例えば，AvePによればISCAS $\geq$ pircsであるが，Q-measureによればpircs $\geq$ ISCASである. これはpircsがISCASよりも高適合文書を優先的に検索できていることを示唆する. 一方，日本語文書検索についてはこのような逆転が見られないことから，際立って高適合文書の検索が得意なシステムはないものと思われる.

次に，各図の中でNTCIR-3と5のグラフを比較すると，NTCIR-5のほうが散らばりが大きく，かつ最高水準が高いことがわかる. ただし，日本語検索に関する水準の差は中国語検索の場合ほど大きくない. 図中には，NTCIR-3と5の両方に参加した前述のチームについてその「進歩」の度合いを点線矢印で示しているが，いずれの場合も矢印は右上がりである. これが本当に「進歩」を意味するか否かについては3.2節で議論する. なお，中国語検索におけるISCASの進歩が著しいが，この主要因は単純なベクトル空間モデルに基づく自前の検索システムからCarnegie Mellon大とMassachusettes大が開発した言語モデルに基づく検索システムLemurに乗り換えたことであると筆者は推測する[5, 14].

†(株)東芝 研究開発センター 知識メディアラボラトリー
tetsuya.sakai@toshiba.co.jp

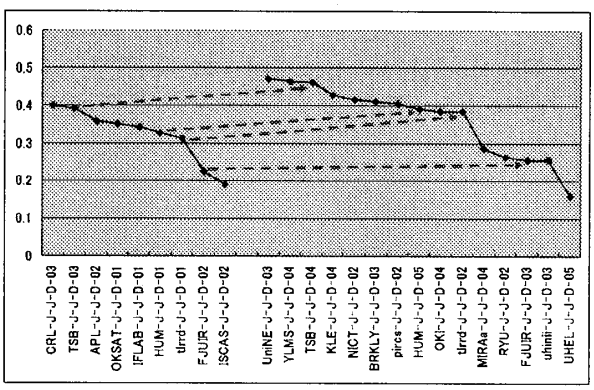

図 1: NTCIR-3(左) と NTCIR-5(右) の日本語単言語 DESCRIPTION runs の比較 (Relaxed AveP)

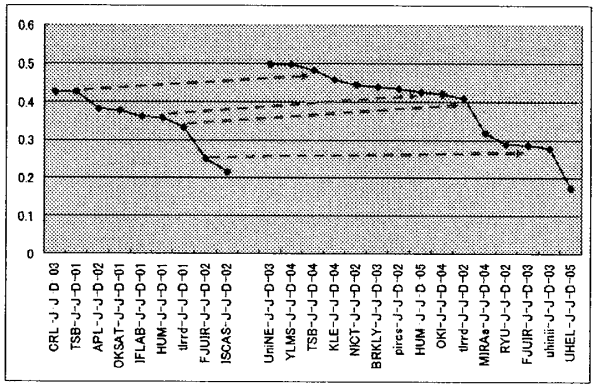

図 2: NTCIR-3(左) と NTCIR-5(右) の日本語単言語 DESCRIPTION runs の比較 (Q-measure)

最新の結果である NTCIR-5 の日本語検索上位チームについて符号検定および (ノンパラメトリック) ブートストラップ検定 [10] を行ったところ，UniNE，YLMS および TSB の間には有意差がなく，UniNE と KLE の間には有意差があった．また，上位 3 チームのうち，UniNE と YLMS は共に形態素解析システム「茶筌」を用いているためか [1, 12]，検索されている文書の顔ぶれが比較的似ていることがわかった．(なお YLMS も前述の Lemur を使用している．) しかし，3 チームとも基本的には Okapi 検索モデル，擬似適合フィードバック，複数検索結果の統合など，英語検索において既に有効性が知られている手法を日本語に適用しており，際立って有効な独自のアプローチは見られないように思える．

上記に比べると，NTCIR-5 の中国語検索の上位チームの結果のほうが「進歩」を予感させる．第 1 位の I2R と第 2 位の UniNE の間には，符号検定では有意差がないがブートストラップ検定では有意差があった．I2R は擬似適合フィードバック前に検索結果を再ランキングする手法を提案しており [4]，このような手法が今後ブレイクスルーにつながる可能性は否定できない．

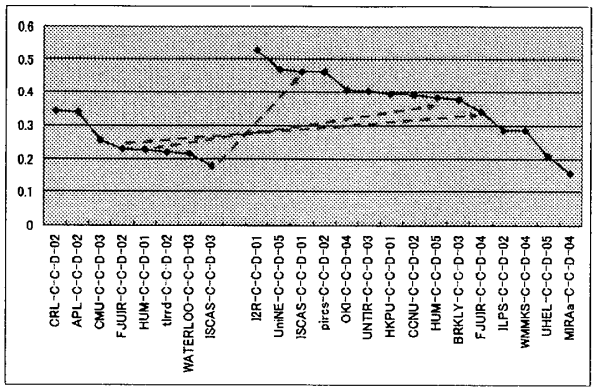

図 3: NTCIR-3(左) と NTCIR-5(右) の中国語単言語 DESCRIPTION runs の比較 (Relaxed AveP)

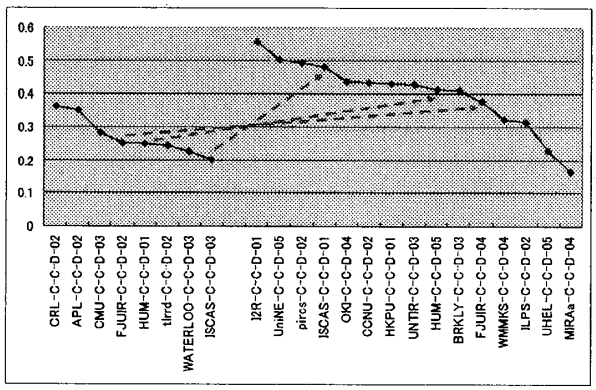

図 4: NTCIR-3(左) と NTCIR-5(右) の中国語単言語 DESCRIPTION runs の比較 (Q-measure)

3.2 NTCIR-3・5 共に参加したチームに着目した考察

表 2 に，図 1〜4 で矢印をつけたチーム，すなわち NTCIR-3・5 共に参加したチームの「進歩」の度合い示す．例えば，HUM は日本語検索において NTCIR-3 では.3272，NTCIR-5 では.3910 の AveP を達成しており，これは+19%の向上である．同様に，NTCIR-3 では.3590，NTCIR-5 では.4246 の Q-measure を達成しており，これは+18%の向上である．さらに同表は，これらの「進歩」に対して対応のないブートストラップ検定 [10] を行った結果も示している．例えば上述の HUM に対する AveP の「進歩」に関する有意確率は 0.229 であり，これは有意ではない．($\alpha = 0.05$, $\alpha = 0.01$ における有意差をそれぞれ $*$, $**$ で示している．)

例えば各日本語検索チームに対する対応のないブートストラップ検定の手順は以下の通りである．

1. NTCIR-3 および 5 の各検索課題に対する検索有効性の値がそれぞれ分布 F および G に従うとしたとき，帰無仮説を $F = G$ とする．
2. NTCIR-3 および 5 の全検索課題 42 + 47 = 89 件からなる集合から，復元抽出により検索課題をランダムに 89 件選出する．うち最初の 42 件を NTCIR-3 分，残り 47 件を NTCIR-5 分のブートストラップ標本とする．以上のリサンプリングを 1000 回試行し 1000 対のブートストラップ標本を生成する．

表 2: NTCIR-3 と 5 の単言語 DESCRIPTION runs の比較 (Relaxed AveP/Q-measure)

team	NTCIR-3	NTCIR-5	%improvement	achieved significance level (unpaired bootstrap hypothesis test)
(a)Japanese monolingual				
FJUIR	.2240/.2502	.2543/.2846	+14%/+14%	0.545/0.509
HUM	.3272/.3590	.3910/.4246	+19%/+18%	0.229/0.222
TSB	.3910/.4246	.4598/.4821	+18%/+14%	0.202/0.257
tlrrd	.3115/.3325	.3827/.4107	+23%/+24%	0.192/0.139
(b)Chinese monolingual				
FJUIR	.2281/.2535	.3425/.3762	+50%/+48%	0.015*/0.007**
HUM	.2257/.2494	.3825/.4135	+69%/+66%	0.004**/0.001**
ISCAS	.1789/.2020	.4632/.4819	+159%/+139%	0.000**/0.000**

表 3: TSB-NTCIR-5to3 と TSB-NTCIR-5 の比較 (同一アルゴリズムで異なるコレクションを検索)

TSB-NTCIR-5to3	TSB-NTCIR-5	%improvement	achieved significance level (unpaired bootstrap hypothesis test)
.4617/.4864	.4560/.4771	-1%/-2%	0.929/0.871

表 4: TSB-NTCIR-3 と TSB-NTCIR-5to3 の比較 (異なるアルゴリズムで同一コレクションを検索)

TSB-NTCIR-3	TSB-NTCIR-5to3	achieved significance level (paired bootstrap hypothesis test)
.3910/.4246	.4617/.4864	0.005**/0.003**

3. 各試行について，ブートストラップ標本の対より検索有効性の平均値の差を求め[‡]，表 2 に示した観測値の差と絶対値の大小を比較する．前者が後者以上である場合はカウンタをインクリメントする．
4. 有意確率の推定値 = カウンタ/1000 とし，この値が有意水準 α より小さい場合に帰無仮説を棄却する．

この結果によれば，NTCIR-3 から 5 への「進歩」は中国語検索においては全て有意であるが，日本語検索においては全て有意でない．言うまでもなく，この「進歩」には，NTCIR-3 と 5 のテストコレクションの難易度自体の違いとシステムの能力の違いの両方が寄与しているはずである．従って，日本語検索の「進歩」に有意差が見られないことに対する説明として，少なくとも以下の 2 つが考えられる．

仮説 A 日本語検索システムは進歩していない．

仮説 B 日本語検索システムは進歩しているが，NTCIR-5 のほうが NTCIR-3 よりもテストコレクションの難易度が高い．

上記いずれの仮説がより現実に近いかを検証するには，両テストコレクションの等価性について議論する必要がある．幸い筆者はチーム TSB に所属しているため，検索アルゴリズムを固定し両コレクションに適用することにより，等価性についての考察がある程度可能である．具体的には，NTCIR-5 に提出した検索結果 (以後，NTCIR-3 の検索結果との混乱を避けるために TSB-NTCIR-5 と呼ぶ) と同じアルゴリズムを NTCIR-3 コレクションに適用した検索結果 (以後，TSB-NTCIR-5to3 と呼ぶ) を新たに作成し，検索有効性の比較を行った．(図 1 と 2 の右側に示した NTCIR-5 の TSB-J-J-D-04 は外部の英語コーパスを利用しているため NTCIR-3 での厳密な再現が難しいので，外部コーパスを利用しない NTCIR-5 の TSB-J-J-D-03 のアルゴリズムを採用した [8]．)

表 3 に TSB-NTCIR-5to3 と TSB-NTCIR-5 の比較結果と，対応のないブートストラップ検定に基づく有意確率を示す．有意確率が 1 に近いことから，NTCIR-3 と 5 のテストコレクションはほぼ等価とみなしてよいと考えられる．(ただしこれは TSB の特定のアルゴリズムに基づく結果であり，厳密には複数のシステムについて同様の検証を行うべきである．また，実際には仮説 B とは逆に NTCIR-3 のほうが難易度が高いという可能性も残る．なぜなら，一般に NTCIR-5 のシステムは NTCIR-3 やこれを包含する NTCIR-4 コレクションに対してチューニングされているからである．) 以上の結果だけを見ると，現実が仮説 B よりも仮説 A に近いように思える．

一方，表 4 は，今度はコレクションのほうを NTCIR-3 に固定し，TSB が NTCIR-3 と 5 でそれぞれ用いたアルゴリズムを比較した結果である．ここで，TSB-NTCIR-3 とは，図 1 の左側に示した NTCIR-3 の TSB-J-J-D-03 のことであり，TSB-NTCIR-5to3 は上述の通り NTCIR-5 のアルゴリズムを NTCIR-3 コレクションに適用したものである．両者は同一の検索課題セットを用いているので対応のあるブートストラップ検定が適用できるが，その結果はいずれも有意差あり ($\alpha = 0.01$) となっている．これはこれまでの議論とは逆に，進化を遂げたチームがあることを示唆する．(ただし，この事例における差の主要因はおそらく TSB が NTCIR-3 の公式結果提出後から Okapi 検索モデルを採用したことである [7, 8]．)

では，表 2(a) で TSB に関する NTCIR-3 と 5 の間の差が有意でなかったのに対し，コレクションを NTCIR-3 に固定した場合の表 4 における差が有意となったのは何故であろうか．それは，対応のないデータ (異なる検索課題セット) に対する検定のほうが，対応のあるデータ (同一の検索課題セット) に対する検定よりも格段に検出力が低いためである．すなわち，表 2(a) のその他のチームについても表 4 と同様にコレクションを固定した検索実験を行えば，統計的有意差を示せる可能性は充分にある．ただ，現在 NTCIR が公開しているデータのみを用いて「検索技術は進歩している」と主張することは難しいということである．まとめると，おそらく NTCIR-3・5 コレクションはほぼ等価であり仮説 B は正しくないが，仮説 A もまた正しくない可能性が残されている．

[‡]この検定方法は，評価値の低い検索課題を重視した評価など，算術平均以外による評価にも適用できる [10]．

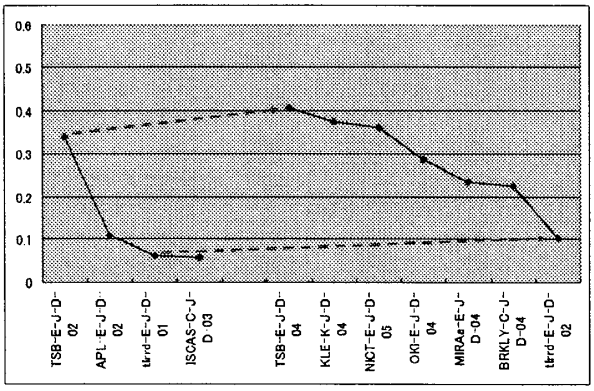

図 5: NTCIR-3(左) と NTCIR-5(右) の言語横断・日本語文書 DESCRIPTION runs の比較 (Relaxed AveP)

4. 言語横断検索の進歩に関する考察

次に，日本語文書を対象とする言語横断検索に主に着目し NTCIR-3 と 5 の結果を比較する．図 5 に，日本語文書を対象とした言語横断検索の成績を検索課題の言語は問わずに AveP により順位づけしたものを示す．(Q-measure による同様の図は割愛する．) このうち NTCIR-3 と 5 の両方に英日検索結果を提出した TSB と tlrrd に着目すると，共に絶対的な検索有効性は NTCIR-5 に対してのほうが若干高いことがわかる．

一方，表 5 は，言語横断検索結果とこれに対応する単言語検索結果を両方提出したチームについて，相対的な検索有効性，すなわち単言語検索の検索有効性に対する割合をまとめたものである§．(参考までに中国語文書を対象とした場合についても掲載している．) 比較的精度が低い tlrrd についてはこの割合が 20%程度から 27%程度に向上しているが，NTCIR-3 および 5 においてトップである TSB については NTCIR-3 において 87%程度，NTCIR-5 において 89%程度とほぼ変化がなく，この結果から言語横断検索特有の技術が格段に進歩していると主張することは難しい．(ただし，TSB は NTCIR-5 において，機械翻訳の第一および第二訳語を同義語グループとして扱い検索に活用しており，その有効性を確認している [7, 8].) すなわち，図 5 において進歩があるように見える主な原因は，単言語検索の検索有効性の値自体が高くなったためである可能性が高い．

5. まとめ

本研究で得られた主な知見は以下のとおりである．

1. NTCIR-3 から 5 までの約 3 年の間に，日本語単言語検索技術が格段に向上したと論ずるに足る証拠はない．ただし，コレクションを固定した実験により進歩を客観的に示せる可能性はある．

2. 同期間において，言語横断検索技術が格段に向上したと論ずるに足る証拠はない．

§相対的な検索有効性は，テストコレクション作成時に検索課題が人手によりいかに翻訳されたかにより値が大きく左右される可能性がある [6]. 例えば日本語テストコレクションの検索課題を人手により英訳する際，「直訳に近い」翻訳を行ったとすると，このデータを用いた英日検索は比較的容易になる．システムにとって日本語の直訳に近い英語を日本語に訳し戻すことが容易であるからである．

表 5: NTCIR-3 と 5 の言語横断検索の検索有効性 (Relaxed AveP/Q-measure)

		absolute	%monolingual
(a)target documents: Japanese			
NTCIR-3	APL-E-J	.1107/.1369	31%/36%
	TSB-E-J	.3404/.3737	87%/88%
	tlrrd-E-J	.0617/.0652	20%/20%
	ISCAS-C-J	.0581/.0649	30%/30%
NTCIR-5	MIRAa-E-J	.2353/.2569	82%/81%
	NICT-E-J	.3601/.3855	87%/87%
	OKI-E-J	.2874/.3193	75%/76%
	TSB-E-J	.4070/.4272	89%/89%
	tlrrd-E-J	.1023/.1165	27%/28%
	BRKLY-C-J	.2253/.2492	55%/57%
	KLE-K-J	.3750/.3960	88%/87%
(b)target documents: Chinese			
NTCIR-3	APL-E-C	.0098/.0094	3%/3%
	ISCAS-J-C	.0563/.0676	31%/33%
NTCIR-5	ISCAS-E-C	.0894/.0900	19%/19%
	OKI-E-C	.0986/.1178	24%/27%
	pircs-E-C	.3235/.3543	70%/72%
	BRKLY-J-C	.1850/.2022	49%/49%

3. 高適合日本語文書の検索が格段に得意なシステムはおそらく今のところ存在しない．

特に知見 1 はあくまで現在入手可能なデータから進歩を証明することが難しいということで，個々のシステムの進歩を否定するものではない．進歩の有無を明らかにするためには，「アルゴリズム固定・新旧コレクション利用」の実験によりコレクションの等価性を検証しつつ，「コレクション固定・新旧アルゴリズム利用」の実験により進歩の度合いを明らかにする仕組みを NTCIR 全体に取り入れるべきであろう¶．

参考文献

[1] Fujita, S.: A Decade after TREC-4 – NTCIR-5 CLIR-J-J Experiments at Yahoo!Japan. *NTCIR-5 Proceedings* (2005)

[2] Harman, D. K.: The TREC Ad Hoc Experiments. *TREC: Experiment and Evaluation in Information Retrieval (Eds: Voorhees, E. M. and Harman, D. K.)*, MIT Press, pp.79-97 (2005)

[3] Kando, N.: Overview of the Fifth NTCIR Workshop. *NTCIR-5 Proceedings* (2005)

[4] Lingpeng, Y. and Donghong, J.: I2R at NTCIR5. *NTCIR-5 Proceedings* (2005)

[5] Min. J. Sun, L. and Zhang, J.: ISCAS in English-Chinese CLIR at NTCIR-5. *NTCIR-5 Proceedings* (2005)

[6] Sakai, T.: Japanese-English Cross-Language Information Retrieval using Machine Translation and Pseudo-Relevance Feedback. *International Journal of Computer Processing of Oriental Languages*, Vol.14, No.2, pp.83-107 (2001)

[7] Sakai, T. *et al.*: Toshiba KIDS at NTCIR-3: Japanese and English-Japanese IR. *NTCIR-3 Proceedings* (2003)

[8] Sakai, T. *et al.*: Toshiba BRIDJE at NTCIR-5: Evaluation using Geometric Means. *NTCIR-5 Proceedings* (2005)

[9] 酒井: よりよい検索システム実現のために: 正解の良し悪しを考慮した情報検索評価の動向. 情報処理, Vol.47, No.2 (2006)

[10] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *ACM SIGIR Proceedings* (2006)

[11] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, to appear (2006)

[12] Abdou, S.and Savoy, J.: Report on CLIR Task for the NTCIR-5 Evaluation Campaign. *NTCIR-5 Proceedings* (2005).

[13] Sparck Jones, K.: Summary Performance Comparisons TREC-2 through TREC-8. *TREC-8 Proceedings* (2000)

[14] Zhang, J. *et al.*: ISCAS at NTCIR-3: Monolingual, Bilingual and Multilingual IR Tasks. *NTCIR-3 Proceedings* (2003)

¶4/17 付の Call for Participation によれば，NTCIR-6 CLIR タスクでは実際にこのような試みが採用される模様である．