機械翻訳を用いた英日・日英言語横断検索に関する一考察

酒井哲也[†] 梶浦正浩[†] 住田一男[†]
Gareth Jones[††] Nigel Collier[†††]

本稿では，日本語テストコレクション BMIR-J2 およびこれを英訳したデータを用い，情報フィルタリングシステム NEAT と機械翻訳システム ASTRANSAC による日・英間の言語横断検索の検索精度評価実験を行う．英語検索要求による日本語文書の検索実験では，文書の日英機械翻訳と検索要求の英日機械翻訳のアプローチを比較し，前者のアプローチにより単言語検索の場合の 90%以上，後者のアプローチにより 80%以上の精度が実現できることを示す．さらに，日本語テストコレクションの検索要求を人手で英訳して英日言語横断検索の模擬実験を行う場合に 2 人の翻訳者を起用することにより，検索精度が人手による翻訳結果に大きく左右されることを示す．一方，日本語検索要求による擬似英語文書の検索実験では，検索要求の日英機械翻訳の前後にローカルフィードバックを行うことにより，単言語検索の場合と同程度の精度を実現する．これは実際の英語文書を検索対象とした場合の精度の上限を示すものと考えられる．最後に，機械翻訳精度と検索精度との関係について考察を行う．

A Study on English-Japanese/Japanese-English Cross-language Information Retrieval Using Machine Translation

Tetsuya Sakai,[†] Masahiro Kajiura,[†] Kazuo Sumita,[†]
Gareth Jones[††] and Nigel Collier[†††]

We study a cross-language IR approach using the NEAT information filtering system and the ASTRANSAC machine translation system. The BMIR-J2 standard Japanese test collection and our own translated data are used for evaluation. In our *English-Japanese* experiments, we compare the retrieval performance of the *document translation* approach and the *query translation* approach, and show that they achieve over 90% and 80% of monolingual performance, respectively. Moreover, in our query translation experiments, we consider the case where the Japanese queries are translated into English by two different translators, and show that the manual translation can greatly affect the performance in traditional evaluation using a monolingual baseline. In our *Japanese-pseudo-English* experiments, we perform local feedback both before and after query translation, and achieve retrieval performance of the monolingual standard. This can be regarded as an upperbound of performance in Japanese-English cross-language IR using real English documents. Finally, we study the relationship between the quality of machine translation and retrieval performance.

1. はじめに

近年のインターネットの普及を背景に，世界中のユーザが自分の使いやすい言語で検索要求を表現し，これに適合する情報を世界中の多様な言語で表現された情報データベースから探し出すという多言語情報検索（multilingual information retrieval: MLIR）の環境が現実的なものとなりつつある．本稿では，MLIR のサブタスクとして，言語 L_1 で記述された検索要求を満たす言語 L_2 ($\neq L_1$) で記述された情報を検索するタスクと定義される言語横断検索（cross-language information retrieval: CLIR）を扱う．CLIR は，たとえば以下の場合に有用である．

- 母国語が L_1 であるユーザ A が，言語 L_2 で記述された情報データベースから情報収集しようとしている．A 氏は，検索要求を言語 L_2 により自分でうまく記述する自信はないが，言語 L_2 で記述された情報を眺めて必要な情報を得る程度の語学

† 株式会社東芝研究開発センター
Toshiba R&D Center
†† Department of Computer Science, University of Exeter
††† 東京大学工学部
Department of Information Science, University of Tokyo

力は有している．

- 母国語が L_1 であるユーザBが，言語 L_2 で記述された情報データベースから情報収集しようとしている．B氏は，言語 L_2 に関する知識をほとんど持たないため，検索された情報をシステムが選択的に言語 L_1 に翻訳してから提示する．

このようなニーズから，欧米ではここ数年CLIRの研究がさかんである．たとえば，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）では，1997年のTREC-6から欧米語の検索要求および文書を対象としたCLIRが扱われるようになった[1]．これにならって，日本でも学術情報センターが日本語検索要求と英語文書を扱うCLIRのタスクを含む評価会NTCIR（NACSIS Test Collection for Information Retrieval Systems）を1999年9月に開催する予定である[2]．

CLIRでは，以下のようなリソースが利用される．

(a) 機械可読辞書（machine-readable dictionaries: MRD）やシソーラス

(b) parallel/comparable/unlinked/monolingual corpora☆

(c) 機械翻訳システム（MT）

また，CLIRにおける検索要求と検索対象の関連性を求める手段，換言すればCLIRを単言語検索（monolingual IR: MIR）に帰着させる手段としては以下が考えられる．

(i) 検索要求を翻訳する．

(ii) 検索対象を翻訳する．

(iii) 言語に依存しない中間言語表現を用いる．

CLIRにおける様々なアプローチにはそれぞれ一長一短がある．たとえば，多言語かつ多方向のCLIRを扱えるという長所を持つ(iii)は，一般に(b)の入手可能性に依存する．また，入手可能性が比較的高い(a)を利用する場合も，多義語の曖昧性解消のために(b)を併用せざるをえない場合が多い．逆に(b)からスタートして(a)の多言語シソーラスを構築し，これをCLIRに利用する事例もある．一方，(c)はもともと曖昧性解消の機構を有しているが，多言語および多方向への拡張のために新たな開発労力を要する．現状では，(a)，(b)と(i)，(iii)の組合せが主流といえる[3)~14)]．しかし，たとえばTREC-6の比較評価では(c)の有効性が再認識されており，さらに(c)のもとでは(i)よりも(ii)のほうが有効であることが示唆されている[1),15)]．(ii)は翻訳コストの観点から敬遠されることが多いが，たとえばユーザの母国語が既知である場合に，前処理の段階で検索対象を翻訳しておくことにより高精度なCLIRを実現するといった応用可能性がある．

本稿では，情報フィルタリングシステムNEAT[16),17)]および機械翻訳システムASTRANSAC[18),19)]を用い，日本語と英語を扱うCLIRのうち(c)と(i)，(ii)の可能性を探るための検索精度評価実験を行う．評価の題材としては，日本語情報検索システム評価用テストコレクションBMIR-J2[20)]☆☆およびこれを独自に英訳したものを用いる．2章で本研究の位置付けについて，3章でCLIRの評価のベースラインとなる日本語単言語検索実験について，4章で英語検索要求による日本語文書の検索実験について，5章で日本語検索要求による英語文書の検索実験について述べる．さらに，6章で考察を述べ，7章でまとめを述べる．

2. 本研究の位置付け

MTをCLIRに利用している従来研究には，主として検索結果の提示のためにMTを利用する試み[14)]や，parallel/comparable corporaの代わりにmonolingual corporaとMTを併用する試み[3),12)]などがある．しかし，前章で述べたように，検索要求あるいは検索対象をMTにより翻訳してCLIRを実現した事例は少ない．さらに，MRDやコーパス利用の研究の中には「現状のMTの性能では高精度なCLIRは期待できない」という主張さえしばしば見られる[9),13)]．一方，MTを積極的に利用している研究者は「現状のMTは人が読んで満足するレベルには達していないが，CLIRには十分有用である」と主張する[21)]．我々もこの立場に立ち，MTの利用により高精度なCLIRが実現できることを示す．

本稿は以下の4つの特徴を持つ．

特徴1：MTを用いた日英間のCLIRを扱い，かつその検索精度をテストコレクションを用いて定量的に評価している．

MTを用いたCLIRで日本語を扱い標準的な検索精

☆ 制約が強い順，したがって実際の入手可能性が低い順に列挙してある．parallel corporaは文書とそれを直接翻訳したものが対応付けられているものであり，comparable corporaは同一トピックについての記述を含む複数言語の文書が対応付けられたものである．対応付けの単位には文書・文・語レベルがあり，対応関係の定義も様々である[3),4)]．これに対しunlinked corporaは対応付けのされていない複数言語のmonolingual corporaの集合である．

☆☆（社）情報処理学会・データベースシステム研究会が，新情報処理開発機構との共同作業により，毎日新聞CD-ROM'94データ版を基に構築した情報検索システム評価用テストコレクションBMIR-J2を利用．

度評価を行った従来研究は，我々の知るかぎり英語検索要求と日本語文書を扱った我々自身の研究[22]のみである．

特徴2：日英間の双方向のCLIRを扱っている．

TREC-6などでは欧米語間の双方向CLIRを扱っている例が見られるが[8),12)]，日本語を扱うCLIRでは従来このような試みは見られなかった．ただし，日本語を扱うCLIRの評価用テストコレクションが現時点では公開されていないため，本稿では，日本語コレクションBMIR-J2の全文書をMTで翻訳することにより，日本語検索要求と擬似英語文書を扱うCLIRの実験を可能にした（本稿ではこれをJe-CLIRと表記する．小文字の“e”は，純粋な自然言語ではなく日英MTの出力であることを表している）．さらに，BMIR-J2の検索要求を人手で英訳することにより，英語検索要求と日本語文書を扱ったCLIRの実験を可能にした（これをEJ-CLIRと表記する）．

特徴3：EJ-CLIRにおいて，検索要求を翻訳した場合と文書を翻訳した場合を比較検討している．

Oardらは，TREC-6の英語検索要求とドイツ語文書を扱ったCLIRにおいて検索要求を英独MTにより翻訳した場合と文書を独英MTにより翻訳した場合の性能比較を行っており，一般には後者のアプローチのほうが多くの文脈を活用できることから高精度を実現する可能性が高いことを示唆している[15)]．しかし，日本語を扱うCLIRにおいてはこのような比較はこれまでに報告されていない．

特徴4：EJ-CLIRにおいて，複数の翻訳者が日本語検索要求を英訳した場合を比較検討している．

通常，言語L_1の検索要求と言語L_2の検索対象を扱うCLIRの実験は以下の手順で行われる．

ステップ1：言語L_2の検索要求Q_2および検索対象D_2を用いてMIRの実験を行う．

ステップ2：Q_2を人手により言語L_1の検索要求Q_1に翻訳する．

ステップ3：Q_1とD_2を用いたCLIRの実験により，MIRの場合の何%の精度が達成できるかを示す．

しかし，上記手順で得られる評価値は，ステップ2の人手による検索要求の翻訳結果に大きく依存すると考えられる．そこで本稿では，2人のバイリンガルにより検索要求を英訳し，それぞれを基にEJ-CLIRの実験を行う．このような試みは欧米でもこれまでに報告されていない．

なお，上記のCLIR評価手順は，文書の言語を揃えてMIRとCLIRの比較を行うというTRECにおいて採用された方式である．一方，NTCIRでは検索対象がparallel corporaに近いため，検索要求の言語を揃えてMIRとCLIRの比較が行われる予定である[2)]．

3. ベースライン：日本語単言語検索（J-MIR）

本章ではCLIRのベースラインとなる日本語MIR（以下，J-MIR）の実験について述べる．我々は文献17)において，確率検索モデルとlocal feedback（以下，LF）を用いた高精度なJ-MIRについて報告した．本稿におけるJ-MIRはこの実験を踏襲したものであるので，概要のみを説明する．なお，検索精度の評価尺度としては11点平均適合率[23)]（以下，11 ptと表記する）を用いる．

(1) BMIR-J2の日本語検索要求50件（Q_Jと呼ぶ）を形態素解析し，名詞などを抽出して初期検索条件を生成する．

(2) 式(1)の検索語重みtwの総和に基づきNEATにより初期検索を行い，パラメータKおよびbを最適化する．

(3) LFを行う．すなわち，初期検索結果の上位n件を正解文書と見なし，この本文中から式(2)の値の降順に新たな検索語（展開語）をm個選定して検索条件に付加する．

(4) 再検索を行いnおよびmを最適化する．なお，本稿ではすべての実験においてNEATの**text**条件のみを利用し，付加する検索条件の条件重みwは0.2で固定した[16),17)]．

$$tw(t,d) = \frac{\log(|C|/df(t)) * tf(t,d) * (K+1)}{K * \left((1-b) + b * \dfrac{l(d) * |C|}{\sum_{d \in C} l(d)}\right) + tf(t,d)} \tag{1}$$

$$rdf(t) * \log \frac{\dfrac{rdf(t)+0.5}{|R|-rdf(t)+0.5}}{\dfrac{df(t)-rdf(t)+0.5}{|C|-df(t)-|R|+rdf(t)+0.5}} \tag{2}$$

ここで，

C：検索対象となる文書集合，

R：正解文書集合，

$l(d)$：文書dの文字数，

$tf(t,d)$：文書d中における検索語tの出現頻度，

$df(t)$：文書集合Cの中で検索語tを含む文書あるいは見出しの数，

K：$tf(t,d)$の影響を調整するために経験的

表1 J-MIRの結果
Table 1 J-MIR results.

名前	Q_J の説明	11 pt
J-INIT	初期検索 ($K = 0.4, b = 0.5$)	0.463
J-LF	LF ($n = 4, m = 15$)	0.491
J-GLF	LFで n, m をグループごとに最適化	0.500

表2 Q_J の初期検索語数に応じて最適化したパラメータ
Table 2 Parameters optimized for Q_J according to the number of initial search terms.

初期検索語数	1	2	3	4	5	6
擬似正解文書数 n	8	6	8	4	4	4
展開語数 m	15	5	15	20	30	5

に決定される定数 ($0 \leq K$),

b : $l(d)$ の影響を調整するために経験的に決定される定数 ($0 \leq b \leq 1$),

$rdf(t)$: 検索語 t を含む正解文書の数.

今回の実験と文献17)との違いは，検索条件展開(query expansion)に加えて検索語の重み調整(term reweighting)も行った点である．これは具体的には，式(1)のlogの項を式(2)のlogの項で置き換えることにより実現される[24]．検索条件展開のみの効果は初期検索精度の5%程度であったが[17]，これと重み調整の併用による効果は6%程度になることを予備実験により確認した．

表1 にJ-MIRの結果を示す．ここで，**J-** はJ-MIRを意味し，**INIT** は K, b を最適化した初期検索結果を，**LF** はLFの n, m を検索要求全体について最適化した結果を表す．また **GLF** は，まず初期検索語数に応じて検索要求をグループ化し，**表2** のように各グループごとに n, m の最適化を行ったものである．これは，LFによりどの程度文脈を補う必要があるかは初期検索条件中にどの程度の情報が含まれているかに依存するという仮説に基づくものである．なお，符号検定[25]で有意差が見られたのは **J-INIT** と **J-GLF** の間 ($\alpha = 0.01$) のみであった．

本稿では，CLIR評価のベースラインとして **J-INIT** および **J-GLF** を用いる．なお，BMIR-J2を用いたJ-MIRにおいてこのような高精度を達成した例はこれまでに報告されていない[17]．また，2章で述べたように，EJ-CLIRのベースラインとしてJ-MIRを用いることは文書の言語を揃えたTREC方式といえるとともに，Je-CLIRのベースラインとしてJ-MIRを用いることは検索要求の言語を揃えたNTCIR方式といえる．**図1** はこの関係を図示したものである．

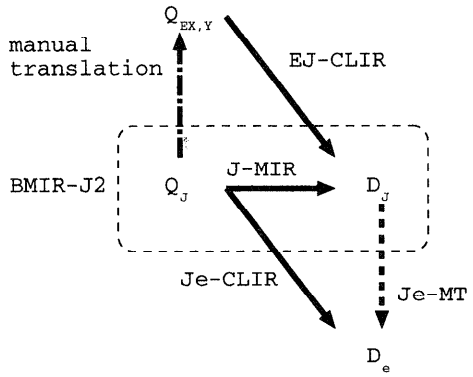


図1 BMIR-J2を用いたJ-MIR, EJ-CLIRおよびJe-CLIR
Fig. 1 J-MIR, EJ-CLIR and Je-CLIR using BMIR-J2.

表3 検索要求文の例
Table 3 Example search requests.

(1) Q_J	行政機関が関係する不況対策
(2) Q_e	Measure against depression relating with a governmental agency
(3) Q_{EX}	Anti-depression measures related to governmental agencies
(4) Q_{EY}	Specific measures taken by government administrators to combat the recession
(5) Q_{jX}	反低下法案は政治の機関に関係があった
(6) Q_{jY}	後退と格闘するために政府行政官によって得られた特定の法案

4. 英語検索要求による日本語文書の検索 (EJ-CLIR)

4.1 EJ-CLIRのための英語検索要求の作成

まず，Q_J をバイリンガルの日本人XおよびYが独自に英訳した．各英語検索要求セットを Q_{EX} および Q_{EY} と呼ぶ．このうちXとYの訳が同一になったものは9件で，これらに関しては以後XとYのすべての検索結果が同一となる．**表3** の(1), (3), (4)に Q_J およびこれに対応する Q_{EX} と Q_{EY} の例を示す．検索要求の表現は翻訳者によりかなりばらつきがあり，これはCLIRの精度に直接影響を与えると考えられる．

4.2 文書の翻訳によるEJ-CLIR

本節では，日本語文書の機械翻訳によるEJ-CLIR(EJDと呼ぶ)の実験について述べる．実験手順は以下のとおりである．

(1) BMIR-J2の文書5,080件 (D_J と呼ぶ) を日英ASTRANSACにより擬似英語文書 (D_e と呼ぶ) に翻訳し，NEATで検索可能にする．なお，本稿のすべての実験において，CLIRの実験の

```
text :1, agenc, depress,
government, measur;
text :0.2, busi, recoveri,
bottom, was, econom, plan,
hang, aspect, cycl,
research, april, consumpt,
postwar, still, low,
monthli, februari, reduct,
adjust, escap, index,
consider, yen, world,
diffusion, recess, real,
late, longest, uptrend;
```

図2 展開後の Q_{EX} の検索条件の例
Fig. 2 An example expanded query from Q_{EX}.

表4 EJD の結果
Table 4 EJD results.

	Q_{EX}		Q_{EY}	
	11 pt	%base	11 pt	%base
EJD-INIT	0.364	79/73	0.323	70/65
EJD-LF	0.416	90/83	0.353	76/71
EJD-GLF	0.427	92/85	0.361	78/72

表5 X と Y の検索要求ごとの比較
Table 5 Per-query comparisons between X and Y.

	(1) EJD-		(2) EJQ-	
11 pt	**INIT**	**GLF**	**INIT**	**GLF**
X > Y	26	27	22	28
X = Y	4	0	10	3
X < Y	20	23	18	19
計	50	50	50	50

ための ASTRANSAC の辞書の変更は英日・日英ともにいっさい行っておらず，専門分野辞書も使用していない．

(2) Q_{EX} および Q_{EY} から生成した初期検索条件を用いて，D_e に対して J-MIR と同様に初期検索および LF を行う．

このように，EJD は EJ-CLIR を英語検索に帰着させる．なお，NEAT では各英単語を stemming して検索を行う．**図2** に，表3の(3)に対応する初期検索条件に対して LF を行った後の検索条件を示す．「text: 1」の後に続いているのが初期検索語，「text: 0.2」の後に続いているのが展開語である[17]．

表4 に EJD の実験結果を示す．**INIT**，**LF**，**GLF** の意味は表1と同様である．Q_{EX} については $K = 0.8$，$b = 0.5$，$n = 10$，$m = 25$，Q_{EY} については $K = 0.6$，$b = 0.7$，$n = 8$，$m = 25$ とした．表中の「%base」の2つの値はそれぞれ11点平均適合率が **J-INIT/J-GLF** の何%であるかを表している．この表の縦方向の対に関する符号検定で有意差が見られたのは，Q_{EX} については **EJD-INIT** と **EJD-LF/EJD-GLF** の間（$\alpha = 0.05$），Q_{EY} については **EJD-INIT** と **EJD-LF** の間（$\alpha = 0.05$），**EJD-INIT** と **EJD-GLF** の間（$\alpha = 0.01$）であった．一方，横方向の対，すなわち Q_{EX} と Q_{EY} の対はすべて有意差なしであった．Q_{EX} と Q_{EY} の平均精度に著しく差があるにもかかわらず有意差が検出されないのは，**表5** の(1)に示したように，ほとんどの検索要求について X が Y より勝っているわけではなく，X が勝っているものと Y が勝っているものがほぼ同数あるためである．実際，第1筆者（翻訳者 X）が Q_{EX} と Q_{EY} の英文を比較した結果，X よりも Y のほうがやや口語的である傾向は見られたものの，前者の翻訳の質が後者より全体的に勝っているとはいえなかった．

TREC などにおける欧米言語どうしを扱う CLIR は，日本語と英語を扱う CLIR よりも一般には容易であると考えられる．たとえば，同一文字セットの欧米言語対においては固有名詞などが同一の表記になることが多いので，翻訳に失敗した語は原語のまま検索語として用いることにより検索に成功する可能性がある．それにもかかわらず TREC-6 における CLIR の精度が単言語検索の 50〜75%程度であったことを考慮すると[1]，我々の EJ-CLIR は高精度であるといえる．とくに，Q_{EX} の **EJD-LF/EJD-GLF** は **J-INIT** の 90%以上の精度を実現している．ただし，Q_{EX} と Q_{EY} の比較から明らかなように，MIR をベースラインとした評価値は翻訳者に大きく依存するため，この数値は1つの目安としてとらえるべきであろう．1人の翻訳者を用いた CLIR の多くの従来研究においても，翻訳者を差し替えることにより MIR をベースラインとした評価値は大きく変わることが予想される．

4.3 検索要求の翻訳による EJ-CLIR

本節では，英語検索要求の機械翻訳による EJ-CLIR（EJQ と呼ぶ）の実験について述べる．実験手順は以下のとおりである．

(1) Q_{EX} および Q_{EY} の自然言語文を英日 ASTRANSAC によりそれぞれ擬似日本語文 Q_{jX} および Q_{jY} に翻訳する．小文字の "j" は，英日 MT の出力であることを表している．なお，Q_J の検索要求が完全に復元されたものは Q_{jX} で 13 件，Q_{jY} で 10 件あった．よって同じ K，b の値を用いればこれらの検索精度は J-MIR と同一（100%）となる．

text :1, 関係, 機関, 政治, 低下, 反, 法案;
text :0.2, 成立, 本会議, 改革, 参院, 否決, 可決, 通過, 与野党, 関連, 審議, 同法案, 野党, 予算, 幹事, 護熙;

図3 展開後の Q_{jX} の検索条件の例
Fig. 3 An example expanded query from Q_{jX}.

表6 EJQの結果
Table 6 EJQ results.

	Q_{jX}		Q_{jY}	
	11 pt	%base	11 pt	%base
EJQ-INIT	0.335	72/67	0.262	57/52
EJQ-LF	0.379	82/76	0.313	68/63
EJQ-GLF	0.387	84/77	0.326	70/65

(2) Q_{jX} および Q_{jY} を基に，D_J に対してJ-MIRと同様に初期検索およびLFを行う．

このように，EJQはEJ-CLIRを日本語検索に帰着させる．表3の(5)，(6)に，同表の(3)，(4)に対応する Q_{jX} および Q_{jY} の例を示す．この例では，“depression”や“recession”といった検索要求中の重要語の誤訳（それぞれ「低下」,「後退」）が見られ，また Q_{jX} の例では“related”を主動詞と解釈したことによる構文解析誤りが見られる．**図3**に，表3の(5)に対応する初期検索条件に対してLFを行った後の検索条件を示す．

表6にEJQの実験結果を示す．Q_{jX} については $K=1.2$, $b=0.3$, $n=8$, $m=15$, Q_{jY} については $K=1.6$, $b=0.2$, $n=6$, $m=25$ とした．この表の縦方向の対に関する符号検定で有意差が見られたのは，Q_{jX} については**EJQ-INIT**と**EJQ-LF/EJQ-GLF**の間（$\alpha=0.01$），Q_{jY} については**EJQ-INIT**と**EJQ-LF**の間（$\alpha=0.05$）および**EJQ-INIT**と**EJQ-GLF**の間（$\alpha=0.01$）であった．また4.2節同様，Q_{jX} のほうが Q_{jY} よりも全体的に精度は高いが，横方向の対はすべて有意差なしであった．なお Q_{jX} と Q_{jY} との検索要求ごとの優劣は表5の(2)のとおりである．

表4と表6の比較から，Oardらの結果[15]同様，EJQよりもEJDのほうが全体的に精度が高いことが分かる．一般に，翻訳時にはEJDのほうが多くの文脈が活用できるため翻訳精度が高くなり，さらに検索時にはEJDのほうが誤訳された語の影響が文書中の他の語により緩和され検索精度が高くなると考えられる．たとえば，翻訳者Xの検索要求ID = 124についてEJQとEJDを比較してみると，EJQの場合，“Regulation relaxation regarding electric communication”（もともとは「電気通信に関する規制緩和」）が「電気的なコミュニケーションに関する規則弛緩」と翻訳されてしまい,「電気的」「コミュニケーション」「規則」「弛緩」が初期検索語となり，初期検索精度は0.000となってしまっている．一方，EJDでは文書全体を翻訳するため，このように訳語が全滅してしまう確率はきわめて低く，対応する初期検索精度は0.262となっている．ただし，比較的小規模なBMIR-J2を用いた今回の実験では，表4と表6で対応する結果の対はすべて有意差なしであった．

なお，MTを用いたCLIRにおいて検索要求翻訳のアプローチと文書翻訳のアプローチを比較する際には，MTの非対称性を考慮する必要がある．我々の実験では，EJDにおいて日英ASTRANSACを，EJQにおいて英日ASTRANSACを用いたが，現状では日英に比べ英日のほうが翻訳品質が良いことが分かっている．たとえば，一般用語辞書の規模を比較すると，英日は約24万語，日英は約14万語である．もしも日英と英日の品質が同程度のMTによりEJDとEJQの比較実験を行えば，BMIR-J2の規模でも有意な差が検出できる可能性はある．

5. 日本語検索要求による擬似英語文書の検索（Je-CLIR）

本章では，日本語検索要求をMTにより英訳し，擬似英語文書を検索するJe-CLIR（JeQと呼ぶ）の実験について述べる☆．これは，1999年9月に開催予定であるNTCIRのJE-CLIRタスクの予備実験として行ったものである．本来ならば予備実験でも英語文書（D_E と呼ぶ）を用いてJE-CLIRの実験を行い，J-MIRやEJ-CLIRとの比較を行うべきである．しかし，我々には D_E の持ち合わせがなく，かつこれまでにNEATの英語文書に対する検索精度評価を行ったことがなかったので，代わりに以下の手順でJe-CLIRの実験を行うことにした．

(1) Q_J を日英MTにより擬似英語文 Q_e に翻訳し，初期検索条件を作成する．
(2) Q_e を基に，4.2節で作成した D_e に対してJ-MIRと同様に初期検索およびLFを行う．

表3の(2)に，同表の(1)に対応する Q_e の例を示す．

上記JeQの検索精度は，検索要求の翻訳によるJE-CLIR（JEQ）の精度の上限と見なすことができる．な

☆ 自然言語の英語文書を用いていないため，文書翻訳によるアプローチ（JeD）を考慮することはできない．

ぜなら，JeQが Q_e による D_e の検索，すなわち同一の日英MTが出力した擬似英語どうしのマッチングを意味するのに対し，JEQは Q_e による自然言語文書 D_E の検索を意味し，一般には前者のほうが容易であると推測されるからである．

さらに本実験では，通常のLFに加えて，我々が日本経済新聞CD-ROM1995年版を利用して独自に作成した別の日本語コレクション TCIR-N1[16),17)]を用いた翻訳前のLF（pre-translation local feedback）[5)]も試みた．CLIRにおける通常のLFが検索要求の翻訳後（post-translation）に検索対象そのものを用いて行うのに対し，翻訳前のLFは，検索要求と言語が同じでかつ検索対象とは独立な monolingual corpus を用いる．すなわち，これは1章で述べた(b)のリソース利用の一例である．コーパスを用いたCLIRにおいて，翻訳前のLFは翻訳精度をあげるための文脈を補う効果がある．

本実験での翻訳前のLFの手順は以下のとおりである．

(1) Q_J によりTCIR-N1の文書5,048件（D'_J と呼ぶ）に対して初期検索を行う．K, b には表1のBMIR-J2における最適値0.4および0.5を流用する．

(2) 通常のLFと同様に，新しい検索語（T'_J と呼ぶ）を抽出する．ただし，簡単のために擬似正解文書数 n' は表1の最適値4で固定し，展開語数 m' のみを最適化する．

(3) T'_J の各語を日英MTにより擬似英語に翻訳し（T'_e と呼ぶ），前述の Q_e に付加する．

図4 にJeQの実験の流れを示す．図中で，e-MIRは擬似英語のMIRを，Je-MTは日英MTを表し，上半分が翻訳前のLFに，下半分が翻訳後のLFに対応する．ここで注意すべきは，コーパス利用のCLIRとは違い，翻訳前のLFは Q_J から Q_e への翻訳精度になんら寄与しないという点である．すなわち，Q_J および T'_J はそれぞれ独立に日英MTにより Q_e および T'_e に翻訳される．また，T'_J は文ではなく語であるので，MTはMRDとしてしか機能しない．

表7 にJeQの結果を示す．Q_e については，$K = 0.2$, $b = 0.9$, $n = 10$, $m = 20$ とした．**PRE** は翻訳前のLFの結果（$m' = 10$），**PRE+GLF** は **GLF** と **PRE** の検索条件をマージした場合の結果を表している．すなわち，図4の上半分の処理のみ施したものが **PRE**，下半分の処理のみ施したものが **LF** および **GLF**，そして両方の処理を施したものが **PRE+GLF** である．符号検定で有意差が見られたのは **JeQ-INIT** と **JeQ-PRE/JeQ-LF** の間（$\alpha = 0.05$）および **JeQ-INIT** と **JeQ-GLF/JeQ-PRE+GLF** の間（$\alpha = 0.01$）であった．全体的に精度が高いのは，前述のとおり D_E の代わりに D_e を用いているためである．Je-CLIRによりMIRと同水準の精度が実現できたことから，もしもASTRANSACの対称性がほぼ成り立つならば，実際のJE-CLIRの精度はEJ-CLIRと同程度になることが予想される．

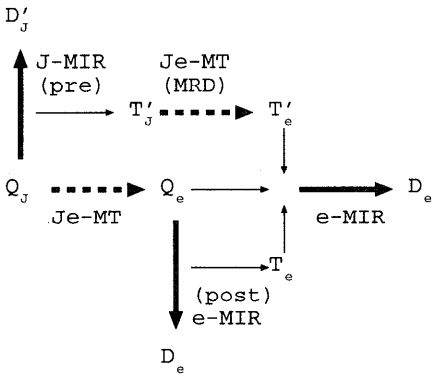


図4 翻訳前後のLFをともなうJeQ
Fig. 4 JeQ with pre/post-translation LF.

表7 JeQの結果
Table 7 JeQ results.

	Q_e	
	11 pt	%base
JeQ-INIT	0.409	88/82
JeQ-PRE	0.426	92/85
JeQ-LF	0.457	99/91
JeQ-GLF	0.461	100/92
JeQ-PRE+GLF	0.462	100/92

JeQ-LF には及ばないものの，**JeQ-PRE** が有効であることは興味深い．翻訳前のLFの効果は，利用するコーパス D'_J が検索対象 D_e とどの程度似た性質を持つか，および展開語の翻訳精度に依存すると考えられるが，本実験の D'_J（TCIR-N1）は1995年の日本経済新聞の記事，D_e は1994年の毎日新聞の記事を日英MTにより翻訳したものであり，D'_j と D_e との関連性はそれほど高くないと考えられる．さらに，前述のとおり，展開語の翻訳は単にMRDの訳語の第1候補を用いていることに相当するため，それほど高精度であるとは考えられない．以上にもかかわらず **JeQ-PRE** が有効であったことは，日本語を扱うCLIRにおいて，またMT利用のCLIRにおいて，翻訳前のLFが有望であることを示していると考えられる．また，翻訳前後のLFの併用 **JeQ-PRE+GLF** と **JeQ-GLF** との間には有意差がなかったが，翻訳前

のLFにおいても n の最適化や **JeQ-GLF** と同様の最適化を行えば，併用の効果が得られる可能性はある．

6. 考　　察

6.1 機械翻訳精度と検索精度の関係について

4章および5章では，従来のCLIR研究の評価方法に従い，単言語検索の精度の何%を実現できるかという総合的な観点から議論を行った．しかし，本研究における検索精度は，機械翻訳システムの翻訳精度に大きく依存していると考えられる．そこで，本節では機械翻訳精度と検索精度の関係について考察を行う．本研究では検索要求翻訳と文書翻訳とを扱っているため，本来は双方の翻訳精度について考慮する必要があるが，文書集合全体に関する翻訳精度を評価することは難しいため，以後，検索要求翻訳に限定して分析を行う．

まず，EJQにおける検索要求の英日機械翻訳結果である Q_{jX} と，JeQにおける検索要求の日英機械翻訳結果である Q_e に対し，第1筆者が以下のような簡単な基準により各検索要求の翻訳精度を3段階で評価した．

ランクA　完璧な翻訳結果であるもの．あるいは，若干言い回しが本来の検索要求と異なっていても，その内容は完全に訳出されているもの．

ランクB　ランクAでもランクCでもないもの．具体的には，検索要求の内容は推定できるが，自然言語として不自然なものがほぼ対応する．

ランクC　検索要求中の主要な概念が明らかに誤訳されているもの，あるいはその翻訳が完全に失敗して原語のまま残ってしまっているもの．あるいは，自然言語として意味不明のもの．

この結果，Q_{jX} についてはランクA＝20件，B＝9件，C＝21件，Q_e についてはランクA＝21件，B＝15件，C＝14件となった．

図5 および **図6** に，それぞれ Q_{jX} および Q_e の各検索要求の初期検索精度をランク別にプロットしたものを示す．横軸は検索要求ID＝101～150であり，縦軸は11点平均適合率である．黒い四角形がランクA，白い四角形がランクB，白い丸がランクCの検索要求を表している．また，**表8** は，Q_{jX} および Q_e についてランク別に初期検索精度の平均値を求めたものである．

図6の解釈は比較的容易である．まず，黒い四角形が図の上のほうに多く，白い丸が下のほうに多く，白い四角形がその間に多い傾向が見られることから，翻訳精度の良し悪しが初期検索精度の良し悪しに直接影響を及ぼす場合が多いことが分かる．このことは表8の Q_e の欄において，ランクが下がるにつれて平均精度が下がっていることからも分かる．さらに，黒い四角形が図の左のほうに多く，白い丸が右のほうに多く，白い四角形がその間に多い傾向が見られるが，これはBMIR-J2の検索要求が構文的・意味的な難易度のほぼ昇順に並べられているため[20]，この難易度の増大にともなって機械翻訳精度が低下したものと考えられる．

図5では，図6ほど翻訳精度と初期検索精度との対応関係が顕著ではない．たとえば，翻訳精度がランク

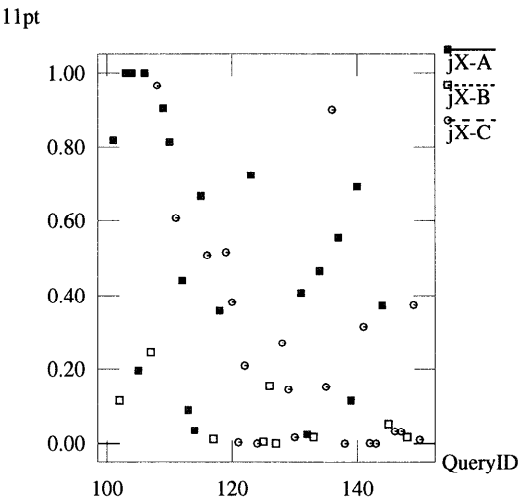


図5　翻訳精度と **EJQ-INIT** の関係（Q_{jX}）
Fig. 5　Relationship between translation accuracy and **EJQ-INIT** (Q_{jX}).

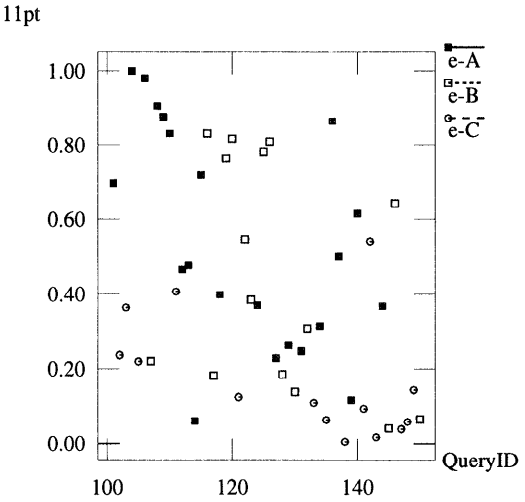


図6　翻訳精度と **JeQ-INIT** の関係（Q_e）
Fig. 6　Relationship between translation accuracy and **JeQ-INIT** (Q_e).

表8 翻訳精度と初期検索精度の平均値の関係
Table 8 Relationship between translation accuracy and mean intial performance.

	EJQ-INIT (Q_{jX})	JeQ-INIT (Q_e)
ランクA	0.534	0.538
ランクB	0.069	0.448
ランクC	0.259	0.173

C（白丸）だが初期検索精度が0.5以上である検索要求が5件も存在し，表8のQ_{jX}の欄において，ランクBの平均精度よりもランクCの平均精度がかえって高くなっている．この第1の原因としては，図5が擬似日本語検索要求Q_{jX}と日本語文書D_Jのマッチングという現実的な検索の実験結果であるのに対して，図6は5章で述べたように擬似英語どうしのマッチングという一種の理想的な検索の実験結果であることが考えられる．すなわち，図6のQ_eの場合でも，もしD_eの代わりに自然言語文書D_Eを用いて実験を行えば，図5と同様にばらつきの多い結果が得られる可能性がある．

第2の原因としては，日本語検索に固有の問題により翻訳精度が初期検索精度に反映されない場合があるということが考えられる．たとえば，図5におけるID = 111（初期検索精度が0.607の白丸）は本来「賃貸住宅」と翻訳されるべきものが「賃貸料用の家」と翻訳され，これは意味をなさないのでランクCと判定されたものである．しかしこの翻訳結果を形態素解析することにより「賃貸」「料」「家」という検索語が抽出され，初期検索はある程度成功している．ここで，仮に「賃貸」ではなく「賃貸料」が形態素解析の結果として得られたとすると，「賃貸住宅」という語は含むが「賃貸料」という語は含まない正解文書が検索できなくなり，検索精度は低下する．このような正解は，実際にブール条件により調べたところ4件あることが分かった．また，4.3節で示したID = 124はこの逆の例で，「電気」でなく「電気的」が検索語として得られてしまったために検索に失敗したものである．以上は，語と語の間の明確な境界がない日本語に対する検索においては，どのような単位を検索語とするかにより検索精度が変わることを示す実例である．このことからも，日本語を扱うCLIRは，欧米言語間のCLIRにはない複雑さをかかえていることがうかがえる☆．

図7は，**JeQ-GLF**により**JeQ-INIT**に対して何%の精度向上が得られたかを図6と対応させてプロットしたものである．黒い四角形，白い四角形，白い丸ともに0%より上にプロットされているものが多いことから，翻訳精度の良し悪しにかかわらず，LFにより精度向上が得られることが多いことが分かる．一方，0%より下にプロットされている検索要求が存在することから，LFが悪影響を及ぼす場合があることが分かる．この点は，ノイズを含むデータからの教師なし学習であるLFの限界であると考えられる．なお，図中の異常値といえるID = 143（向上率が700%以上の白丸）は，検索要求「経営陣刷新」の英訳が失敗し，「刷新」という語が原語のまま残ってしまった場合を表している．欧米言語間の検索とは違い，文字セットの異なる原語は検索にまったく役立たないので，初期検索語としては「経営」の訳語である“management”をstemmingした“manag”のみが用いられた．このため，初期検索精度は0.017と非常に低かったが，LFにより検索精度は0.143まで向上している．このように，MTの失敗は文字セットの異なる言語間の検索において致命的であるが，その影響をLFが緩和できる場合があることが分かる．

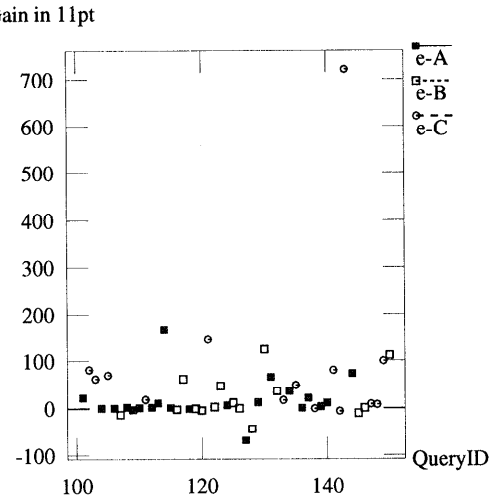


図7 **JeQ-GLF**の**JeQ-INIT**に対する向上率（Q_e）
Fig. 7 Percentage gain of **JeQ-GLF** over **JeQ-INIT** (Q_e).

6.2 LFの効果の符号検定結果について

本研究では，J-MIRにおいては**J-INIT**と**J-GLF**の間のみに有意差が見られたが，EJ-CLIRにおいてはXとYともに**EJD-INIT**と**EJD-LF/EJD-GLF**の間および**EJQ-INIT**と**EJQ-LF/EJD-GLF**の間に有意差が見られ，Je-CLIRにおいては**JeQ-INIT**と**JeQ-PRE/JeQ-LF/JeQ-GLF/JeQ-PRE+GLF**の間に有意差が見られた．

☆ さらに，4.3節のID = 124の例は，“communication”という語を「通信」と訳すか「コミュニケーション」というカタカナ外来語に訳すかという困難さを示す例でもある．

すなわち，LFの効果は，MIRよりもCLIRにおいてより顕著であった．これは主に，MIRよりもCLIRの初期検索精度が平均的に低いことに起因すると考えられる．我々はこれまでに，文献17)における検索要求グループ[20]別の評価により，初期検索精度がある程度高い場合には，LFによる精度向上の余地がなくなる傾向があることを示したが，今回の実験でも同様の現象が起こったものと考えられる．

7. まとめ

本稿では，機械翻訳システムとローカルフィードバックを利用した日・英間のCLIRの検索精度評価を行い，その有効性を示した．英語検索要求による日本語文書の検索実験では，文書翻訳のアプローチによりMIRの90%以上の精度を，検索要求翻訳のアプローチによりMIRの80%以上の精度を実現し，またこれらの評価値は検索要求の翻訳者に依存することを示した．日本語検索要求による擬似英語文書の検索実験では，検索要求の翻訳前後のローカルフィードバックを扱い，MIRと同程度の精度を実現した．これは，日本語検索要求と英語文書を扱うCLIRにおける我々のアプローチの上限を示すものと考えられる．いずれの実験においても，ローカルフィードバックが検索精度向上のための有効な手段であることが分かった．さらに，機械翻訳精度と検索精度の関係について，具体例を見ながら考察を行った．

図8は，**J-GLF**，**JeQ-PRE+GLF**，**EJD-GLF**（Q_{EX}），**EJQ-GLF**（Q_{jX}）の再現率–適合率曲線[23]である．また**表9**は，これらのCLIRの結果を検索要求ごとに**J-GLF**と比較した結果である．この表から明らかなように，全体の3分の1程度の検索要求についてはCLIRのほうがMIRよりもかえって精度が高い．これは2章で述べた下記の主張を裏付けるものであると考えられる．

「現状のMTは人が読んで満足するレベルには達していないが，CLIRには十分有用である」

たとえば，表3の(2)は，翻訳結果としてはいま一歩であるが，これを初期検索条件に変換すると，表3の(3)の人手で作成した検索要求から作成した初期検索条件と同一になる．すなわち，初期検索語 “agenc, depress, goverment, measur” を含む図2の1番目の条件に変換される．現状のMIRの性能自体が必ずしも十分であるわけではないが，少なくともMTとCLIRとの親和性は高いといえる．

CLIRではMIRに増して検索結果の提示方法が重要となる．たとえば，1章で述べたB氏のようなユーザに対しては，言語障壁を隠蔽するユーザインタフェースを考慮する必要があると考えられる[11),14)]．今後は，検索精度・翻訳精度の向上にとどまらず，ユーザが検索結果から実際に情報を入手するまでのインタラクションを含めた研究を進める予定である．

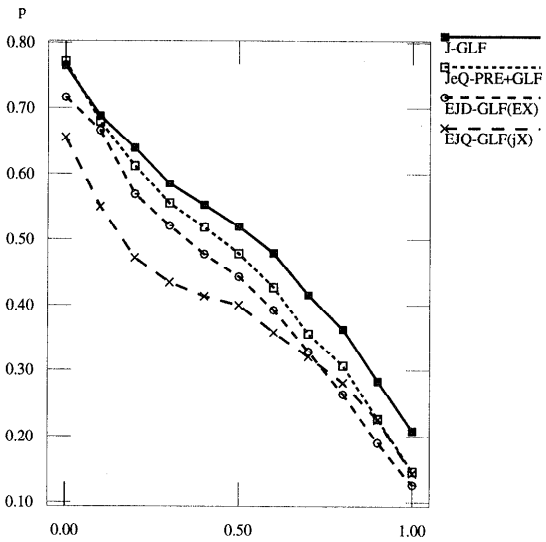


図8 再現率・適合率曲線
Fig. 8 Recall/precision curves.

表9 **J-GLF**との検索要求ごとの比較
Table 9 Per-query comparisons with **J-GLF**.

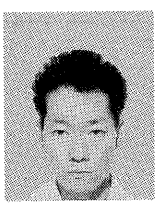
J-GLFと比べ11 ptが	(1) **EJD-GLF**		(2) **EJQ-GLF**		(3) **JeQ-PRE+GLF**
	Q_{EX}	Q_{EY}	Q_{jX}	Q_{jY}	Q_e
高い	13	11	16	13	16
同じ	0	1	2	1	0
低い	37	38	32	36	34
計	50	50	50	50	50

参考文献

1) Schauble, P., et al.: Cross-Language Information Retrieval (CLIR) Track Overview, TREC6, `http://trec.nist.gov/pubs/trec6/t6_proceedings.html` (1997).
2) NTCIR: `http://www.rd.nacsis.ac.jp/~ntcadm`
3) Mateev, B., et al.: ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval, TREC6, `http://trec.nist.gov/pubs/trec6/t6_proceedings.html` (1997).
4) Sheridan, P., et al.: Experiments in Multi-

lingual Information Retrieval using the SPIDER system, *Proc. ACM SIGIR '96*, pp. 58–65 (1996).
5) Ballesteros, L., et al.: Resolving Ambiguity for Cross-language Retrieval, *Proc. ACM SIGIR '98*, pp.64–71 (1998).
6) Carbonell, J.G., et al.: Translingual Information Retrieval: A Comparative Evaluation, IJCAI'97, pp.708–714 (1997).
7) 藤井ほか：多言語情報検索のための複合語翻訳，情処学会情報学基礎研究会，98-FI-51, pp.55–62 (1998).
8) Gaussier, E., et al.: Xerox TREC-6 Site Report: Cross Language Text Retrieval, TREC6, `http://trec.nist.gov/pubs/trec6/t6_proceedings.html` (1997).
9) Hull, D., et al.: Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval, *Proc. ACM SIGIR '96*, pp.49–57 (1996).
10) Kando, N., et al.: Cross-lingual Information Retrieval using Automatically Generated Multilingual Keyword Clusters, *Proc. 3rd International Workshop on Information Retrieval with Asian Languages*, pp.86–94 (1998).
11) 菊井ほか：インターネット情報ナビゲーションにおける多言語機能，情処学会自然言語処理の応用に関するシンポジウム，`http://titan.mcnet.ne.jp/titan-help/jp/3.html` (1995).
12) Littman, M.L., et al.: Automatic Cross-language Information Retrieval using Latent Semantic Indexing, *Cross-Language Information Retrieval*, Grefenstette, G. (Ed.), pp.51–62, Kluwer Academic Publishers, Norwell, Massachusetts (1998).
13) Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-language Information Retrieval, *Proc. ACM SIGIR '98*, pp.55–63 (1998).
14) Yamabana, K., et al.: A Language Conversion Front-End for Cross-language Information Retrieval, *Cross-Language Information Retrieval* Grefenstette, G. (Ed.), pp.93–104, Kluwer Academic Publishers, Norwell, Massachusetts (1998).
15) Oard, D.W., et al.: Document Translation for Cross-Language Text Retrieval at the University of Maryland, TREC6, `http://trec.nist.gov/pubs/trec6/t6_proceedings.html` (1997).
16) 酒井ほか：情報フィルタリングのためのブール式と文書構造を利用した検索条件生成と検索精度評価，情報処理学会論文誌，Vol.39, No.11, pp.3076–3083 (1998).
17) 酒井ほか：確率モデルに基づく日本語情報フィルタリングにおけるフィードバックによる検索条件展開および検索精度評価，情報処理学会論文誌，Vol.40, No.5, pp.2429–2438 (1999).
18) Hirakawa, H., et al: EJ/JE Machine Translation System ASTRANSAC – Extensions towards Personalization, *Proc. MT Summit III*, pp.73–80 (1991).
19) Kinoshita, S., et al.: ASTRANSAC – Toshiba Machine Translation System, *Proc. MT Summit VI*, p.284 (1997).
20) 酒井ほか：情報検索システム評価のためのテストコレクション，*Computer Today*, Vol.9, No.87, pp.31–35, サイエンス社 (1998).
21) Gachot, D.A., et al.: The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Cross-Language Information Retrieval, *Cross-Language Information Retrieval*, Grefenstette, G. (Ed.), pp.105–118, Kluwer Academic Publishers, Norwell, Massachusetts (1998).
22) Jones, G., et al.: Cross-Language Information Access: a case study for English and Japanese, 情処学会情報学基礎研究会，98-FI-51, pp.47–54 (1998).
23) Witten, I.H., et al.: *Managing Gigabytes: Compressing and Indexing Documents and Images*, pp.148–151, Van Nostrand Reinhold, (1994).
24) Robertson, S.E., et al.: *Simple, Proven Approaches to Text Retrieval*, Computer Laboratory, University of Cambridge (1994).
25) Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments, *Proc. ACM SIGIR '93*, pp.329–338 (1993).

(平成 10 年 12 月 17 日受付)
(平成 11 年 9 月 2 日採録)

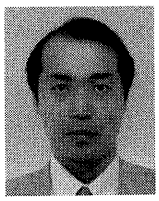


酒井 哲也（正会員）

昭和 43 年生．平成 5 年早稲田大学大学院理工学研究科工業経営学専門分野修士課程修了．同年（株）東芝入社．自然言語処理，情報検索・フィルタリングの研究開発に従事．

梶浦 正浩（正会員）

昭和41年生．平成3年慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．平成6年同専攻後期博士課程単位取得退学．同年（株）東芝入社．情報検索・情報フィルタリング，ECの研究開発に従事．

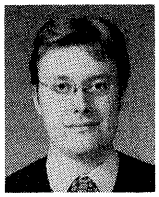


住田 一男（正会員）

昭和32年生．昭和57年東京工業大学大学院総合理工学研究科物理情報工学専攻修士課程修了．同年（株）東芝入社．自然言語処理，情報検索・フィルタリングの研究開発に従事．工学博士．

Gareth Jones

Gareth Jones received a B.Eng in Electrical and Electronic Engineering from the University of Bristol, UK in 1989, and a PhD in 1994 from the same institution for his thesis examining the application of linguistic models in automatic speech recognition. From 1993 to 1996 he was a Research Associate at the University of Cambridge, UK carrying out research on multimedia information retrieval. In 1996 he was appointed as a Lecturer in Computer Science at the University of Exeter, UK. From 1997 to 1998 he was a Toshiba Fellow at the Toshiba R&D Center. His research interests include: information retrieval, information access, speech recognition, natural language engineering, and virtual technology. He is a member of the UK IEE and British Computer Society IRSG.



Nigel Collier

Nigel Collier received his BSc. in computer science from Leeds University in the UK in 1992, an MSc. in machine translation and a Ph.D. in language engineering from UMIST in the UK in 1994 and 1996 respectively. From 1990 to 1991 he was a visitor at the University of Tokyo. From 1996 to 1998 he was a Toshiba Fellow at Toshiba R&D Center and is currently a JSPS research associate at Tokyo University investigating information extraction from genome domain texts. His current research interests are in machine translation, cross-language information retrieval, and knowledge acquisition. He is a member of the ACM and SIGART as well as the British Computer Society.