日英言語横断検索における関連性の重ね合わせモデルの効果

金沢輝一[†] 相澤彰子[††]
高須淳宏[††] 安達淳[††]

筆者らは自然言語の持つ意味曖昧性による情報検索の精度低下の問題に対して関連性の重ね合わせモデルによる検索を提案してきた．本論文では，提案手法が言語横断検索において，単一言語検索時と同様に検索精度を向上させることを情報検索のためのテストセット NTCIR 1，2 を用いた評価実験により示す．提案手法である関連性の重ね合わせモデルは，著者キーワードなどの情報に基づいて文書をクラスタリングすることで，索引語の重要度計算を tf·idf などの手法より高い精度で行うものである．本論文の実験ではその効果の言語独立性を示すために，辞書を用いた翻訳手法あるいはコーパスに基づく翻訳手法と組み合わせてテストセット上で言語横断検索の精度を比較，特性を分析する．実験の結果は，提案手法が言語横断検索でも単一言語検索時と同程度の 4〜9% の精度向上を示し，また問合せ拡張（query expansion）と組み合わせることでより高い検索精度が得られた．

Effectiveness of the Relevance-based Superimposition Model for Cross-language Information Retrieval

Teruhito Kanazawa,[†] Akiko Aizawa,[††] Atsuhiro Takasu[††]
and Jun Adachi[††]

We have proposed Relevance-based Superimposition (RS) model for IR which is expected to solve the problem of semantic ambiguity. In this paper, we show the effectiveness of our proposed IR method in cross-language information retrieval by the experiments with NTCIR 1, 2 multilingual IR test sets. The proposed RS model modifies the document feature vectors using document clusters organized according to the relevance of documents, and it is expected to achieve higher precision of retrieval, independent of language. The results of our experiments with dictionary- and corpus-based translation methods indicate that our document feature modification model achieves 4 to 9% improvement, avoiding the difficulties of language- or domain-dependent parameters. Furthermore, the combination of our model and query expansion achieves higher improvement, which is more than the summation of their individual effectiveness.

1. はじめに

World Wide Web をはじめさまざまな場面で言語横断検索に対する要求が高まっている．検索対象の記述言語と異なる言語で問合せを行う言語横断検索では翻訳過程で生じる自然言語の意味曖昧性が検索精度の低下を引き起こす要因となっている．筆者らは情報検索における自然言語の意味的曖昧性の問題に着目し，関連性の重ね合わせモデル（RS モデル）による検索を提案してきた[1)~3)]．これは，ベクトル空間モデルの情報検索において，同一キーワードを含むなどの関連性に基づいて文書をクラスタリングし，関連文書に含まれている索引語の情報を用いて文書ベクトルを補正するもので，学術論文データベースのように同一の概念がさまざまな筆者によって多様な表現で表される場合に特に効果を発揮する．本論文では RS モデルの言語横断検索における適用可能性を評価するため，NTCIR 1，2 テストセットを用いた言語横断検索（日本語問合せによる英語文書の検索）を行い，和英の単一言語検索と精度，特性を比較する．その際，複数の翻訳手法と組み合わせることで翻訳によって生じる意味曖昧性に対する本手法の寄与の安定性を検証する．

本論文は以下の構成となっている．まず 2 章で意味的曖昧性の問題と RS モデルについて説明する．次に 3 章で言語横断検索について述べ，4 章で評価実験の概要と結果を報告する．そして実験結果に基づいた考

† 東京大学大学院工学系研究科
Graduate School of Engineering, University of Tokyo
†† 国立情報学研究所
NII（National Institute of Informatics）

察を 5 章で行い，最後にまとめを述べる．

2. RS モデル

2.1 情報検索における意味曖昧性の問題

検索対象の文献と問合せ表現はともに自然言語の意味的曖昧性を持っている．すなわち同表記異義によって問合せとは無関係な文書が検索されたり，同義異表記によって検索されるべき文書が検索できなかったりする場合がある．意味的曖昧性による検索精度の低下を抑えることは情報検索の最も重要な課題の 1 つであり，これまで多くの研究がなされてきた．それらは大きく 3 つに分類できる．すなわち，query expansion（以下 QE）など問合せ表現を補正するもの，特異値分解などにより文書の特徴空間を変化させるもの，文書の特徴量を補正するもの，の 3 つである．

QE は検索者の入力した問合せ表現に関連する語句を加えることで問合せの特徴ベクトルを拡張するものである．問合せは情報量が比較的少ないため，これに基づいて検索者の意図を汲み取り適切な語句だけを自動的に加えることは困難である．実用上は検索対象のデータベースに合わせてパラメータの調整などを行う必要がある[4]．

特徴空間を変化させるアプローチは Latent Semantic Index[5] などの手法に代表されるように，特異値分解によって索引語を単位ベクトルとする特徴空間から概念を単位ベクトルとする低次元の特徴空間に射影することで意味のマッチングを行おうというものである．これらの手法の課題は特異値分解の計算コストが他手法に比べて非常に大きいことであり，大規模のデータベースに対する適用に向けて研究が進められている．

文書の特徴量を補正するアプローチは，問合せ表現より多くの情報を用いて意味的曖昧性に対処する．これにより，問合せによっては検索精度が極端に低下するという，QE に発生しがちな現象を回避することができると考えられる．Salton は学術文献の引用関係を用いて，引用文献の説明文を被引用文献の索引語に用いる手法を提案している[6]．近年では hypertext のリンク情報を用いて，リンクの説明を被リンク文書の特徴量解析に用いることが文書分類の分野で提案されている[7),8]．これらは検索あるいは分類の精度向上に寄与することが様々な評価実験によって示されており，Web 検索などへの応用も期待されている．課題としてあげられているのは，引用あるいはリンクの説明は必ずしも被参照文書の概要を表現するキーワードではないが，有効な情報のみを抽出するような手法がドメインに依存しない一般的な形で確立していない点である．また引用やリンクの関係は参照関係が偏っている場合が多く，被参照回数の少ない文書は適切な補正を受けられない可能性がある．筆者らの提案している**関連性の重ね合わせモデル**（Relevance-based Superimposition モデル，以下 RS モデル）は著者キーワードなどの情報に基づいて文書をクラスタリングし，これを解析することで文書の特徴ベクトルを補正するというものである．ここでいう文書クラスタは従来のクラスタリングに基づいた検索手法群における排他型のものではなく，1 つの文書が複数のクラスタに属することを許している．排他型クラスタリングでは，たとえば「ニューラルネットワークを用いた画像処理」に関する文書は「ニューラルネットワーク」か「画像処理」のいずれか一方の話題にのみ分類され，もう一方との関連性を表現することができないという問題があった．提案手法では 1 つの文書が複数の話題に属しているという，より自然なモデルを表現できる．以下に RS モデルの詳細を述べる．

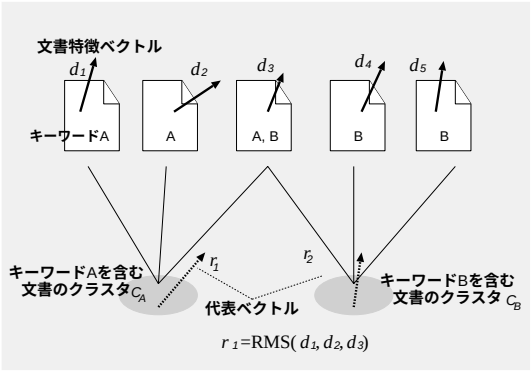

図 1 代表ベクトルの生成
Fig. 1 Representative vector generation.

2.2 非排他型文書クラスタの生成

文書群 $\{\mathrm{d}_1, \mathrm{d}_1, \cdots, \mathrm{d}_N\}$ で構成されたデータベースを仮定する．また，各々の文書に対応する文書ベクトルを $\{\boldsymbol{d}_1, \boldsymbol{d}_1, \cdots, \boldsymbol{d}_N\}$ と定義する．RS モデルでは文書を非排他型クラスタ $\{C_1, C_2, \cdots, C_M\}$ に分類する．今回の実験ではクラスタは文書から抽出したキーワードによって形成されている．たとえば**図 1** のようにデータベース中にキーワード A と B の 2 つのキーワードが存在した場合，キーワード A を含む文書はクラスタ C_A に，キーワード B を含む文書はクラスタ C_B に属する．また，キーワード A，B をともに含む文書は C_A と C_B の両方に属するものとする．

2.3 代表ベクトルの生成

RS モデルによる文書ベクトルの拡張は，クラスタの代表ベクトル生成と，代表ベクトルを用いての文書

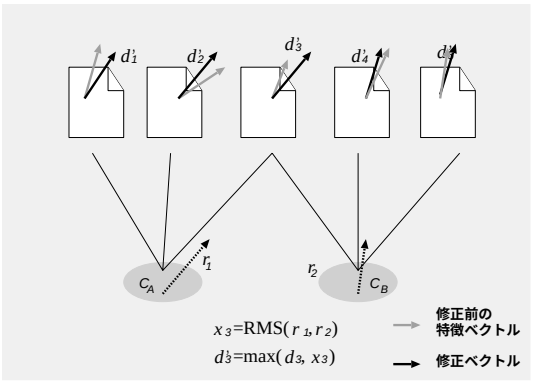

図 2 文書ベクトルの補正
Fig. 2 Document vector modification.

ベクトルの実質的な修正の 2 つの段階を経て行われる．

まず最初の段階として，文書クラスタごとに代表となる特徴ベクトルを生成する．このベクトルは文書ベクトルと同じ特徴空間内のベクトルであり，同数の次元を持つ．モデルとしては特徴空間を構成する特徴量に制限はないが，実験では tf·idf の考え方に基づく統計量（後述）を用いている．クラスタ C の代表ベクトル $\boldsymbol{r}$ は C に属する全文書のベクトルを引数とする代表ベクトル生成関数によって生成される．α-関数族[9] から派生するいくつかの関数の評価[10] によると，最も良い性能を示す代表ベクトル生成関数は，Root-Mean-Square を用いたもので，代表ベクトル $\boldsymbol{r}$ の第 i 要素 r_i を次のように求める関数である．

$$r_i \equiv \sqrt{\frac{1}{|C|}\sum_{\mathrm{d}_j \in C} {d_{j,i}}^2} \tag{1}$$

ただし，$|C|$ はクラスタ C に含まれる文書数，$d_{j,i}$ は文書 d_j のベクトル $\boldsymbol{d}_j$ の第 i 要素である．

2.4 文書ベクトルの補正

次に，代表ベクトルを用いて各文書のベクトルを拡張する．文章が属するすべての文書群の代表ベクトルの Root-Mean-Square と，文書ベクトルとを要素ごとに比較して，前者が大きければ文書ベクトルの新たな要素として置き換える（**図 2**）．

$$d'_{j,i} \equiv \max(d_{j,i}, x_{j,i}), \tag{2}$$

$$x_{j,i} \equiv \sqrt{\frac{1}{k_j}\sum_{l=1}^{k_j} {r_{l,i}}^2} \tag{3}$$

ただし，$r_{1,i}, \cdots, r_{k_j,i}$ は文書 d_j が属す文書群 $\boldsymbol{r}_1$, $\cdots$, $\boldsymbol{r}_{k_j}$ の代表ベクトルの第 i 要素である．これらの関数は文書が持っていた固有の特徴量が重ね合わせによって過度に平均化されないものを評価実験[10] をもとに選択した．

3. 言語横断検索

3.1 手法の分類

言語横断検索では問合せと検索対象文書との言語の違いを翻訳などによって吸収する必要がある．その手法は，問合せの翻訳，文書の翻訳，特徴空間の変換の 3 つに大別できる．検索対象の文書を翻訳するアプローチはコストの面で実用的とはいえず，特徴空間の変換も 2 章で述べたように計算量の問題がある．以上の理由から問合せ表現を翻訳するアプローチが一般的である．翻訳のための手法は，機械翻訳，辞書による翻訳，コーパスを用いる翻訳に分類できる．機械翻訳は文脈を汲み取ることで辞書による逐語翻訳以上の品質を得ることができるが，情報検索の問合せはきわめて短い表現であり，文脈を自動的に認識するのに十分な情報量を持っているとはいえない．すなわち問合せの翻訳に機械翻訳を用いることは適当ではないと考えられる．一方，辞書やコーパスを用いた言語横断検索は単一言語検索と比しても十分実効的な検索精度を達成している[11)~13)]．辞書による学術文書の翻訳では，文書の著者によって作り出された新語など辞書に記述されていない語句の取扱いが課題となっており，コーパスなどからの対訳語句自動抽出が注目されている．本論文の評価実験では EDR の日英対訳辞書を用いた翻訳と，NTCIR コーパスから抽出した対訳関係を用いた翻訳を行った．

3.2 辞書を用いた翻訳

EDR 日英対訳辞書[14] は日本電子化辞書研究所によって構築された電子化辞書で，実験で使用した Version 1.6 には 364,431 個の日本語語彙に対応する英語訳を収録している．

今回の実験では以下の手順を繰り返すことで問合せを翻訳した．

(1) 日本語の問合せ表現の先頭から，最長一致の表現を辞書から探す．
(2) 対応する対訳表現のうち第一義をもって翻訳する．

3.3 コーパスを用いた翻訳

NTCIR コーパスは文献単位の対応情報を持っている対訳コーパスである．付与されている著者キーワードはキーワード単位の言語対応をとることが要求されているわけではないが，**図 3** の例のように大部分は対応関係にある．この点に着目して対訳関係を抽出することで翻訳を行うことができる．ただしすべてのキーワードが対訳関係にあるわけではない点に留意する必要がある．予備実験では無作為に選んだ 1,000 対のう

ち 93%が意味的に対応していた[15]．

筆者らはグラフ理論に基づいて誤った関係を切り離すことでキーワードをクラスタリングし，類義語辞書を生成する手法を提案しており，Kando らの言語横断検索手法の評価にも用いられている[16),17)]．今回の実験では対訳の選択方法として

- どの集合から選択するか，
- 単一の候補を選ぶか複数選ぶか，
- 複数選ぶ場合，重み付けを行うか，

を基準に以下の 6 通りに設定した．

(1) 直接の対応関係がある語句から，最大頻度の 1 つを選択する．
(2) 直接の対応関係がある語句すべてを対訳として用い，重み付けはしない．
(3) 直接の対応関係がある語句すべてを対訳として用い，QE の語句と同様の pseudo relevance feedback（後述）によって重み付けを行う．
(4) キーワードクラスタから，最大頻度の 1 つを選択する．
(5) キーワードクラスタの語句すべてを対訳として用い，重み付けはしない．
(6) キーワードクラスタの語句すべてを対訳として用い，QE の語句と同様の pseudo relevance feedback によって重み付けを行う．

EDR でも複数の対訳が存在した場合にそれらをすべて用いることは可能だが，EDR は類義語辞書ではないために専門用語的な語彙に対しては複数の対訳が存在しない場合も多いこと（**図 4**）から今回の実験では行わなかった．なおキーワードをクラスタリングする場合のパラメータ[15] は $N_\alpha = 3$，$N_\beta = 3$，$N_\varepsilon = 10$ とした．このとき，キーワードクラスタは 271,437 個に分割された．最大のクラスタは 5,558 対（日本語 236 語，英語 557 語）のキーワードを含んでいた．

EDR あるいはコーパスから対訳を抽出した例を図 4 に示す．

注意しなければならないのは検索対象と対訳抽出のためのコーパスが同一である場合と異なる場合で語彙の網羅率が変化して検索精度に影響する可能性がある点である．実験では対訳の抽出は NTCIR 1 コーパス（後述）からのみ行い，検索対象を NTCIR 2 コーパスのみとした場合と NTCIR 1，2 コーパスの両方とした場合を分析した．後者は NTCIR 2 ワークショップでの標準的な言語横断タスクに相当する．

日本語： 遺伝アルゴリズム / 最適探索 / 学習
English: Genetic Algorithms / Optimization / Machine Learning

図 3 NTCIR コーパスの著者キーワードの例
Fig. 3 An example of keyword lists given by the authors of documents.

4. 評価実験

4.1 概　　要

4.1.1 テストコレクション NTCIR

ここでは評価実験で用いた情報検索のためのテストコレクション NTCIR について概要を述べる．詳細は文献 18)，19) を参照されたい．

文書コーパスは国内の学会発表の抄録と科学研究費補助金研究成果概要から構成されており，言語などによって**表 1** の 6 つに分類されている．ntc1-j1 と ntc1-e1，ntc2-j1g と ntc2-e1g，ntc2-j1k と ntc2-e1k はそれぞれ一部が文献単位の対訳関係を持っている．以下，ntc1 の接頭辞を持つデータ群を NTCIR 1 コーパス，ntc2 の接頭辞を持つデータ群を NTCIR 2 コーパスと呼ぶことにする．

検索課題は 1 文程度の自然文質問であり，NTCIR 1 コーパスに対してのみ正解判定が行われている問合せ#1〜#83 と，NTCIR 1，2 両方のコーパスに対して正解判定が行われている#101〜#149 が存在する．

EDR による対訳：information retrieval（全 1 語彙）
コーパス中の対応関係にある英語表現（括弧内は頻度）： information retrieval (546), information search (11), retrieval (10), information retrival (9), information retrieral (6), ..., internet (1), query processing (1)（全 73 語彙）
キーワードクラスタを形成した英語表現： information retrieval (572), search (165), retrieval (140), information retrieval, ..., codebook search (6)（全 238 語彙）
pseudo relevance feedback に基づく QE によって「information retrieval」に対して補われた英語表現： keyword, document, query, text, word, media, ...

図 4 キーワード「情報検索」の対訳リストと query expansion の結果
Fig. 4 Translated/expanded keyword lists for "*information retrieval*".

表 1 NTCIR コーパス
Table 1 The NTCIR corpus.

	言語	コーパス名	文書数
学会発表抄録	日本語	ntc1-j1	332,918
		ntc2-j1g	187,080
	英語	ntc1-e1	116,177
		ntc2-e1g	77,433
科研費研究成果概要	日本語	ntc2-j1k	287,063
	英語	ntc2-e1k	57,545

表 2 コーパスから抽出されたキーワードの数
Table 2 The numbers of keywords extracted from the corpora.

	キーワード数	5 件以上の文書に出現する数
NTCIR 日本語	851,218	90,761
NTCIR 英語	632,930	46,418

4.1.2 評価方法

評価実験は問合せ#101〜#149 を用いて，NTCIR の標準的な評価方法に従い，問合せごとの最大上位 1,000 件における再現率と適合率を求めた[19]．以下，J-J タスクとは日本語問合せによる日本語文書の検索，E-E タスクとは英語問合せによる英語文書の検索，J-E タスクとは日本語問合せによる英語文書の検索をさす．

表 2 に，NTCIR コーパスから抽出して RS モデルの文書クラスタ作成に用いたキーワードの数を示す．含んでいる文書の数があまりに小さいクラスタは個々の文書ベクトル固有の特徴が代表ベクトル作成時に過大に反映されてしまいノイズとなるので，文書数 4 以下のクラスタは除外した．

4.2 検索システム R^2D^2

筆者らは RS モデルの評価のために文献検索システム R^2D^2 （RetRieval system for Digital Documents）を作成した[3]．これはベクトル空間モデル[20] に基づいており，RS モデルや QE，翻訳などの処理の組合せを適宜変更することで様々な条件の検索精度を評価することができるもので，NTCIR 1，2 ワークショップにおいて baseline としての十分な検索精度を有していることを確認している[21]．本論文における評価実験は R^2D^2 を用いて，tf·idf の考え方に基づく統計量のみによる検索結果（baseline）と RS モデルを適用した場合の結果について精度比較を行う．

4.2.1 形態素解析

索引語，検索語について日本語の形態素解析は chasen version 1.51[22] を用い，自立語を抽出，語幹レベルで索引を作成した．英語は空白などのデリミタによって単語を切り出し，ストップワードを取り除いた後，Porter の語幹切り出しアルゴリズム[23] と不規則変化動詞の辞書を併用して語幹のレベルで索引を作成した．和英どちらの場合も単語単位の解析のみ行い，句の認識は行わなかった．

4.2.2 検索語の重み付け

R^2D^2 では検索語 $\{q_0, q_1, \ldots, q_k\}$ からなる問合せ $\boldsymbol{Q}$ に対する文書 d_j の検索語 q_i の重みを 3 つの特徴量：

- 文書中での語の出現頻度（term frequency）に基づく特徴量：$f_T(j,i)$
- 全文書中で語を含む文書の数（document frequency）に基づく特徴量：$f_D(i)$
- 語の共起頻度（term cooccurrence）に基づく特徴量：$f_C(i,\boldsymbol{Q})$

によって定義する．

baseline における f_T は，NTCIR，TREC テストセットを用いた予備実験[24] で最も良い性能であった，

$$f_{T0}(j,i) \equiv \frac{1}{\pi}\arctan\left(\alpha\frac{tf_{j,i}}{F(j)}+\beta\right)+0.5 \quad (4)$$

を NTCIR 1，2 に対する最適値である $\alpha = 100$，$\beta = -0.5$，$F(j) = \sum_i tf_{j,i}$ という条件で用いた．また式 (1)〜(3) に $d_{j,i} = f_{T0}(j,i)$ を代入して得られた $d'_{j,i}$ を RS モデルを適用した場合の f_T とした．

$f_D(i)$ には，索引語 t_i を含む文書数を df_i，全文書の数を N としたときの $f_D(i)$ として，予備実験で性能の良かった，

$$f_D(i) \equiv \log(N/df_i) \quad (5)$$

を用いた．

$f_C(i,\boldsymbol{Q})$ は，文書 d_j に出現する検索語の種類を c_j，検索語 t_i が出現する文書の集合を Δ_i として，

$$c(i) \equiv \sum_{\mathrm{d}_j\in\Delta_i}\sum_{\mathrm{t}_k\in\mathrm{d}_j} f_D(k) \quad (6)$$

$$\overline{c}(i) \equiv \sum_{\mathrm{d}_j\notin\Delta_i}\sum_{\mathrm{t}_k\in\mathrm{d}_j} f_D(k) \quad (7)$$

$$f_C(i,\boldsymbol{Q}) \equiv \log\frac{c(i)}{df_i} - \log\frac{\overline{c}(i)}{N-df_i} \quad (8)$$

を用いた．式 (8) では，問合せの話題に関連度の高い文書集合における情報量（右辺第 1 項）と，補集合における情報量（右辺第 2 項）との差分をとっている．

そして，検索語 q_i の重みを

$$w(i,j,\boldsymbol{Q}) \equiv f_T(j,i)\cdot f_D(i)\cdot f_C(i,\boldsymbol{Q}) \quad (9)$$

と定義し，文書 d_j の得点は $\sum_{i=0}^{k} w(i,j,\boldsymbol{Q})$ とする．f_T と f_D は問合せに依存しない特徴量であり，$(f_T(j,0)\cdot f_D(0),\ldots,f_T(j,k)\cdot f_D(k))$ を文書ベクトルと見なして RS モデルを適用した．

表 3 問合せの翻訳の誤り分類
Table 3 Types of query translation.

分類	EDR（実験番号 1，5）		NTCIR（実験番号 2，6）		NTCIR（実験番号 9）	
A	88	(33%)	125	(52%)	149	(54%)
B	17	(6%)	15	(6%)	33	(12%)
C	44	(16%)	20	(8%)	26	(9%)
D	23	(9%)	6	(3%)	8	(3%)
E	27	(10%)	33	(14%)	21	(8%)
F	44	(16%)	35	(15%)	37	(13%)
G	26	(10%)	4	(2%)	4	(1%)
計	269		239		278	

分類は，NTCIR 2E の対応する問合せ表現（Q_m）と翻訳結果（Q_a）の関係に基づいて行っている．

(A) Q_m と Q_a が同表現．
(B) Q_m と Q_a は一部一致する類似表現．
(C) Q_m と Q_a は類義表現．
(D) Q_a は Q_m の関連表現だが，より抽象的．
(E) Q_a は Q_m には存在しない表現．
(F) Q_m には存在するが，Q_a には含まれない表現．
(G) まったく異なる意味の表現．

総数が手法によって異なるのは Q_m あるいは Q_a の一方にしかないものなどがあるため．

4.2.3 Query Expansion

2 章で述べたように，RS モデルは自然言語の曖昧性の問題に対する 3 つのアプローチのうち文書の特徴量を補正するものである．我々は評価実験によって RS モデルと QE の性質の差異を比較する．また，文書の特徴量補正と問合せの補正は排他的なものではなく，組み合わせることによってより高い精度の検索を行えるものと考えられるので，この点も実験によって検証する．

実験では pseudo relevance feedback に基づく自動 QE[23] を評価用システムに実装して用いた．補われる語は初期検索の結果で上位 D 件の文書に含まれる索引語のうち，tf·idf の平均が大きい T 語を検索語に補う．すなわち文書の得点，$\mathrm{score}(\mathrm{d}_j) = \sum_{i=0}^{k} w(i, j, \boldsymbol{Q})$ の大きい D 件の文書集合を $\mathcal{D}$ として，索引語ごとに

$$\sum_{\mathrm{d}_j \in \mathcal{D}} w(i, j, \boldsymbol{Q}) \tag{10}$$

を求め，その値の大きい T 語を含めた検索語集合 $\boldsymbol{Q}' = \{q_0, q_1, \ldots, q_k, q_{k+1}, \ldots, q_{k+T}\}$ に対する文書 d_j の得点，$\sum_{i=0}^{k+T} w(i, j, \boldsymbol{Q}')$ を算出する．言語横断検索では，翻訳した問合せに対して同様の操作を行った．NTCIR の問合せ#31〜#83 ならびに TREC 3 によるパラメータチューニングにおいて，D と T の最適値は $D = 30$，$T = 10$ であった．実験では NTCIR の他の問合せを用いて，パラメータと QE の性能の関係を調べる．

4.3 翻訳精度の分析

まず翻訳の精度について分析を行う．**表 3** は問合せの翻訳誤りを分類したものである．分類は各手法による翻訳結果と NTCIR の E-E タスク用問合せとの比較に基づいて行った．A，B，C は意味的に同等である表現に翻訳されたものであり，A は完全に一致したもの，B は表記的に一部一致する類義表現，C は表記的には異なるが類義表現であるものである．EDR の日英対訳辞書を用いた場合は全体の約 55%，NTCIR コーパスから抽出した対訳関係を用いた場合は約 66%，対訳関係をクラスタリングした場合は約 75%がこれらに含まれる．一方，翻訳誤りと分類される D〜G の事例を調べると，EDR では医学や数学などの分野の語彙，たとえば「虚血性心疾患」「コアグラーゼ」「有限要素法」など，問合せの重要な概念を翻訳できなかったのに対して，コーパスに基づく翻訳手法では「各種抗菌物質の MRSA に対する効果について」という問合せ中の「各種（various）」のように検索の精度には大きな影響を持たない語句の割合が多いことが分かった．また，コーパスから抽出した対訳をクラスタリングした場合，対訳が発見できなかった分類 E の割合が特に減少し，類似表現に翻訳された分類 B の割合が高まっている．これはコーパス上で直接の対訳関係を持たない類似表現がクラスタリングによって発見されたことを示す結果である．

4.4 検索実験の結果

表 4 は各手法の平均適合率，**表 5** は同じ翻訳条件で baseline とその他の検索手法の平均適合率の比，**表 6** は同じコーパスに対する英語の単一言語検索（実験番号 1〜4 に対しては 16，5〜11 に対しては 15）と言語横断の場合の比を示している．実験番号は問合せ，検索対象コーパス，翻訳の種類などの組合せごとに番号が与えられており，詳細は表 4 に示すが実験番号 1〜11 が言語横断検索，12〜16 が単一言語検索，1〜4 は検索対象と翻訳辞書の情報源が異なる場合，5〜11

表 4 各手法の平均適合率
Table 4 Average precisions.

実験番号	検索対象	問合せ	翻訳辞書	対訳語選別	対訳語重み付け	baseline	QE	RS	QE+RS
1	NTCIR 2E	NTCIR 2J	EDR	第一義	—	.1212	.1275	.1275	.1379
2			NTCIR 1	最大頻度の対訳	—	.1805	.1852	.1893	.1921
3			NTCIR 1	すべての対訳	あり	.1846	—	.1968	—
4			NTCIR 1	クラスタ内すべて	あり	.1832	—	.1967	—
5	NTCIR 1E+2E	NTCIR 2J	EDR	第一義	—	.1641	.1846	.1716	.1931
6			NTCIR 1	最大頻度の対訳	—	.2369	.2476	.2522	.2653
7			NTCIR 1	すべての対訳	なし	.1224	.1247	.1287	—
8			NTCIR 1	すべての対訳	あり	.2451	—	.2635	—
9			NTCIR 1	クラスタ内最大頻度	—	.2401	.2441	.2508	.2574
10			NTCIR 1	クラスタ内すべて	なし	.1524	.1565	.1517	—
11			NTCIR 1	クラスタ内すべて	あり	.2441	—	.2650	—
12	NTCIR 1J	NTCIR 1J				.3059	.3270	—	—
13	TREC SJM	TREC 3				.2318	.2578	—	—
14	NTCIR 1J+2J	NTCIR 2J				.2841	.2886	.3020	.3103
15	NTCIR 1E+2E	NTCIR 2E				.2984	.3044	.3160	.3226
16	NTCIR 2E	NTCIR 2E				.2329	.2382	.2515	.2552

表 5 baseline に対する各手法の精度向上
Table 5 Precision improvements over baseline.

実験番号	QE	RS	QE+RS
1	1.0520	1.0520	**1.1378**
2	1.0260	1.0488	1.0643
3	—	1.0661	—
4	—	1.0737	—
5	1.1249	1.0457	**1.1767**
6	1.0452	1.0646	**1.1199**
7	1.0188	1.0515	—
8	—	1.0751	—
9	1.0167	1.0446	**1.0721**
10	1.0269	0.9954	—
11	—	1.0856	—
12	1.0690	—	—
13	1.1122	—	—
14	1.0158	1.0630	**1.0922**
15	1.0201	1.0590	**1.0811**
16	1.0228	1.0799	1.0957

太字は QE と RS を併用した場合の寄与が，QE と RS の寄与の単純和よりも大きい場合．

表 6 E-E タスクに対する各手法の精度
Table 6 Precisions of cross-lingual runs over monolingual runs.

実験番号	baseline	QE	RS	QE+RS
1	.5204	.5353	.5070	.5404
2	.7750	.7775	.7527	.8248
3	.7926	—	.7825	—
4	.7866	—	.7821	—
5	.5499	.6064	.5430	.5986
6	.7939	.8134	.7981	.8224
7	.4102	.4097	.4072	—
8	.8214	—	.8339	—
9	.8046	.8019	.7937	7979
10	.5107	.5141	.4801	—
11	.8180	—	.8386	—

は検索対象の一部が翻訳辞書の情報源となっている場合である．実験番号 15 は NTCIR 2 標準の E-E タスク，14 は NTCIR 2 標準 J-J タスク，12 は NTCIR 1 標準 J-J タスクである．12 と 13 は QE のパラメータチューニングに用いた．実験 15 における baseline の平均適合率は 0.2984 で，QE は 2%，RS モデルは 6%の精度向上を達成した．QE の最適パラメータは $D = 40$，$T = 10$ であった．この値はチューニングで得られた最適値 $D = 30$，$T = 10$ と異なるが，平均適合率の違いは 0.3%程度であった．

翻訳手法に関しては 1 と 2 あるいは 5 と 6 の比較から静的な辞書を用いるよりもコーパスから抽出した対訳関係を用いた場合の方が高い検索精度を得られることが分かる．一方，コーパスを用いる場合にはすべての対訳候補を重み付けする手法の精度が高く，次いで最大頻度の 1 つを選ぶ手法，すべての候補を重み付けせずに用いる手法の順となった．キーワード対をクラスタリングしたことの検索精度に対する効果は，7 と 10 のように誤り候補の影響が大きい場合には強く現れたが，重み付けをしたり（8 と 11 の比較）最大頻度の語を選んだりした場合（6 と 9）ではこれらの処理も誤り候補の影響を抑える作用を持っているのでクラスタリングの寄与が大きく現れなかった．

J-J，E-E タスクにおいて RS モデルと QE を組み合わせた場合の精度向上率は，各々の手法を単体で適用した場合の向上分の単純な合計を上回っていたが，言語横断検索においても同様の傾向であった．

図 5 は各手法の特徴を問合せごとの平均適合率についてのヒストグラムで表したものである．RS モデルの寄与は QE よりも正方向に偏っており，問合せによ

らず検索精度を向上させることを示している．一方，QE は問合せによっては検索精度を低下させることが分かる．また表 5 でも QE は 2〜12%と効果のばらつきが大きく，また言語横断検索では翻訳精度が高い場合ほど効果が小さくなるなど QE のパラメータチューニングの困難さを端的に示す結果となったのに対し，RS モデルは実験 10 を除いて 4〜9%の範囲で安定した効果を得られた．TREC，NTCIR などの大規模テストセットを用いた評価で 5〜10%の平均適合率の変化は顕著な差であるとされており[25]，RS モデルは十分な効果を実現していると考えられる．

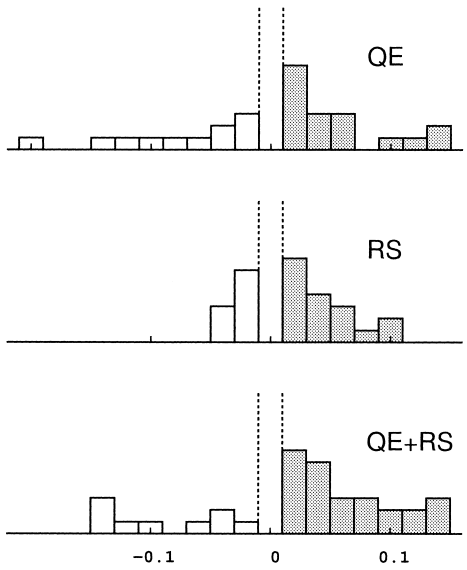

図 5 問合せごとの寄与のヒストグラム（実験番号 9 における比較）
Fig. 5 Histogram of precision improvement over baseline per query.

5. 考　　察

RS モデルは QE と組み合わせることで相乗効果を持つことが示された．言語横断検索において，複数の対訳を用いて翻訳を行うことは，問合せの持つ意味曖昧性に対して query expansion と同種のアプローチで対処しているとも考えられ，もし RS モデルが QE と相乗効果を持つならば複数の対訳を選択した場合にも RS モデルの効果が増しているはずである．表 5 の実験 5 と 7，9 と 11 の RS モデルの精度向上率を比較すると確かにその傾向が現れている．すなわち，検索対象のベクトル補正である RS モデルの作用は問合せ表現の補正とは異なる性質を持ち，相補的に検索精度を向上させるといえる．

表 7 は言語横断検索において RS モデルが著しく検索精度を下げた例である．正解文書に付与されているキーワードは綴りの間違いを含むため，正しく文書クラスタリングが行われていない．その結果，ベクトル補正も適切に行われず相対的に順位を下げてしまっている．現在の実装では出現文書数が 5 に満たないキーワードは RS モデルに用いられず，そのようなキーワードが全体の約 9 割に達する．大半は綴りの間違いであり，問合せの翻訳で用いたキーワードのクラスタリング手法を適用することで誤った綴りのキーワードによる文書クラスタを正しい綴りのキーワードによ

表 7 コーパスを用いた翻訳に対する RS モデルの失敗例 Q102：糖尿病治療のための異種膵島移植の長期生着例,（正解数 3）
Table 7 Inappropriate case for RS model.

順位	RS 得点	タイトルと著者キーワード
4 → 7	1.1830	Title: Direct ultraviolet irradiation of fetal pancreas *zenografts* (xenografts) for reversal of the experimental diabetic dogs KWDs: Fetal pancreas / ***Panceas*** (Pancreas) **transplantation** / Xenograft / Ultraviolet Irradiation / Diabetes / Dog / Immunoalteration
5 → 8	1.2541	Title: Development of bioartificial endocrine pancreas using glucose-responsive B-cell line Min6 KWDs: diabetes / **bioartificial endocrine pancreas** / glucose transporter / MIN6 cell / insulin secretion
7 → 20	0.0000	Title: Xenogeneic Tnukplantation of Islets Employing Bioartificial Pancreas KWDs: Bioartificial *Pamrens*(Pancreas) / MRPB / Three-*layr*(layer) Microbead / B Cell line / MIN6 / HINS51 / ***prcine Isiets***(Porcine Islets) / Device Change
(16 → 4)	6.0123	Title: Macroencapsulated artificial pancreas KWDs: **Islet transplantation** / **Artificial pancreas**
(8 → 5)	4.3392	Title: Antigen modulation on pancreatic islet transplantation KWDs: **Islet transplantation** / Antigen modulation / gamma-irradiation / Anti-lymphocyte serum

順位は baseline → RS．不正解文書は順位に括弧を付けて区別した．
太字は問合せに関連したキーワード，太字斜体は関連するはずだが綴りの誤りを含むもの．参考のため正しい表現を（キーワード）として示した．RS 得点とは baseline における各文章の得点と RS モデルによって補正が行われたベクトルによる得点の差．

る文書クラスタに統合することができ，これによってRS モデルの性能を高められるのではないかと考えている．

今後の課題としては，著者キーワードが付与されていないデータベースへの対応があげられる．RS モデルは文書クラスタの作成のために良質なキーワードのリストが必要であり，それがない場合には自動キーワード抽出を行うか，あるいはキーワード以外の情報から文書関連性を解析する必要がある．いずれにしても文書クラスタの性質が変わることが予想されるので，検索への影響を調べることが重要である．

なお，今回の実験ではキーワードのクラスタリングの意図した効果が他の処理の効果に埋もれてしまう現象が見られたが，キーワードのクラスタリングは誤対応を切るだけでなく直接対応関係にない類義的な語を候補に含められる可能性があり，これは QE と同一の作用と考えられる．類義語を多く含むようにクラスタリングのパラメータを変えた場合の検索精度への寄与を分析する必要があろう．

6. おわりに

本論文は NTCIR テストセットを用いた実験結果により，RS モデルが言語に対して独立の効果を持つことを示した．

RS モデルは文書の非排他的なクラスタを作成し，これを用いて文書の特徴ベクトルを補正するもので，単一言語検索で 6%，言語横断検索で 4〜9%の精度向上を達成した．言語横断検索においては問合せの翻訳に辞書を用いた場合，コーパスから対訳関係を抽出して用いた場合の双方で効果を示した．また，RS モデルを query expansion と組み合わせることで，より大きな効果を得ることができた．表 4 が示すように，2 つの手法を統合した場合には，それぞれを単体で適用した場合の単純な和を上回る検索精度の向上率が得られている．すなわち，query expansion が適切に問合せを補正することで RS モデルの性能を高めていると思われる．この効果は言語横断検索において対訳表現を複数選択することで問合せ側の意味曖昧性に対処した場合にも見られた．

謝辞 本研究は日本学術振興会未来開拓事業 JSPS-RFTF96P00602 の援助を受けている．本研究は「国立情報学研究所共同研究員規程」に基づく共同研究として，国立情報学研究所（NII）の構築した情報検索システム評価用テストコレクション NTCIR-1 および NTCIR-2（本格版研究目的使用）を使用した．このテストコレクションには，`http://research.nii.ac.jp/ntcir/acknowledge/thanks1-ja.html` のリストに示されている学協会によって開催された学会における発表論文の要旨，および，文部省科学研究費補助金研究成果の概要が含まれている．また，日本電子化辞書研究所が構築した EDR 電子化辞書を使用した．

参考文献

1) Kanazawa, T.: R^2D^2 at NTCIR: Using the Relevance-based Superimposition Model, *Proc. NTCIR Workshop 1*, Tokyo, pp.83–88 (1999).
2) Kanazawa, T., Takasu, A. and Adachi, J.: A Relevance-based Superimposition Model for Effective Information Retrieval, *IEICE Transactions*, Vol.E83-D, No.12, pp.2152–2160 (2000).
3) Kanazawa, T., Takasu, A. and Adachi, J.: R^2D^2 at NTCIR 2 Ad-hoc Task: Relevance-based Superimposition Model for IR, *Proc. NTCIR Workshop 2*, Tokyo, pp.204–210 (2001).
4) Mitra, M., Singhal, A. and Buckley, C.: Improving Automatic Query Expansion, *SIGIR '98*, pp.206–214 (1998).
5) Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A.: Indexing by Latent Semantic Analysis, *J. American Society for Information Science*, Vol.41, No.6, pp.391–407 (1990).
6) Salton, G.: Associative document retrieval techniques using bibliographic information, *J. ACM*, Vol.10, No.4, pp.440–457 (1963).
7) Chakrabarti, S., Dom, B. and Indyk, P.: Enhanced hypertext categorization using hyperlinks, *SIGMOD '98*, pp.307–318 (1998).
8) Attardi, G., Gullì, A. and Sebastiani, F.: Automatic Web Page Categorization by Link and Context Analysis, *Proc. THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pp.105–119 (1999).
9) 林 幸雄：個人選考による情報アクセスに適したデータモデルについて，情報処理学会データベースワークショップ '98（情報処理学会研究報告），98-DBS-116(2), Vol.98, No.58, pp.381–388 (1998).
10) 金沢輝一，高須淳宏，安達 淳：関連性の重ね合わせモデルによる文書検索，電子情報通信学会データ工学ワークショップ '99（電子情報通信学会研究報告），Vol.99, 鹿児島 (1999).
11) Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval, *SIGIR '98*, pp.55–63 (1998).
12) Ballesteros, L. and Croft, W.B.: Resolving

Ambiguity for Cross-Language Retrieval, *SIGIR '98*, pp.64–71 (1998).
13) Fujita, S.: Notes on the Limits of CLIR Effectiveness: NTCIR-2 Evaluation Experiments at Justsystem, *Proc. NTCIR Workshop 2*, Tokyo, pp.181–188 (2001).
14) 日本電子化辞書研究所：EDR. `http://www.iijnet.or.jp/edr/`
15) 相澤彰子，影浦　峡：学術文献の和英著者キーワードを用いた類義語クラスタの自動生成，情報処理学会論文誌，Vol.41, No.4, pp.1180–1191 (2000).
16) Aizawa, A. and Kageura, K.: An Approach to the Automatic Generation of Multilingual Keyword Clusters, *Proc. COMPTERM'98*, pp.8–14 (1998).
17) Kando, N. and Aizawa, A.: Cross-Lingual Information Retrieval using Automatically Generated Multilingual Keyword Clusters, *IRAL'98*, Singapore (1998).
18) Kando, N.: Overview of the Second NTCIR Workshop, *Proc. NTCIR Workshop 2*, pp.35–44 (2001).
19) NTCIR: `http://research.nii.ac.jp/ntcir/`
20) Salton, G. and McGill, M.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
21) 江口浩二，栗山和子，神門典子：テストコレクションにおける検索課題の難易度予測への挑戦，情報処理学会研究報告，2001-FI-63, Vol.2001, No.74, pp.17–24 (2001).
22) 奈良先端科学技術大学院大学自然言語処理学講座：日本語形態素解析器 'ChaSen'. `http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html`
23) Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison-Wesley (1999).
24) 金沢輝一，高須淳宏，安達　淳：英語テキストにおける関連性の重ね合わせモデルの検索特性，情報処理学会データベースワークショップ 2000（情報処理学会研究報告），2000-DBS-122, Vol.2000, No.69, 岩手，pp.57–64 (2000).
25) Voorhees, E.: Variations in Relevance Judgements and the Measure of Retrieval Effectivenss, *SIGIR '98*, pp.315–323 (1998).

(平成 13 年 9 月 25 日受付)
(平成 13 年 10 月 31 日採録)

（担当編集委員　藤原 譲）

金沢　輝一（学生会員）
1997 年東京大学工学部卒業．1999 年同大学大学院工学系研究科修士課程修了．現在同大学院工学系研究科博士課程に在籍．

相澤　彰子（正会員）
1985 年東京大学工学部卒業．1990 年同大学大学院工学系研究科博士課程修了．工学博士．1990 年から 1992 年，イリノイ大学アーバナ・シャンペイン校客員研究員．現在，国立情報学研究所助教授．自動用語抽出，統計的情報処理，遺伝的アルゴリズム等の研究に従事．電子情報通信学会，人工知能学会，言語処理学会，IEEE 各会員．

高須　淳宏（正会員）
1984 年東京大学工学部卒業．1989 年同大学大学院工学系研究科博士課程修了．工学博士．同年学術情報センター助手．学術情報センター助教授を経て現在国立情報学研究所助教授．データベースシステム，文書画像処理，機械学習に関する研究に従事．電子情報通信学会，人工知能学会，ACM，IEEE 各会員．

安達　　淳（正会員）
1981 年東京大学大学院工学系研究科博士課程修了．工学博士．東京大学大型計算機センター助手，文部省学術情報センター研究開発部助教授，教授を経て現在国立情報学研究所教授．オンライン情報システム，分散処理システム，情報検索，電子図書館システム等の開発に従事．電子情報通信学会，ACM，IEEE 各会員．