情 報 学 基 礎 63－10
（２００１．７．２５）

# 関連性の重ね合わせモデルを用いた日英言語横断検索

金沢 輝一† 相澤 彰子‡ 高須 淳宏‡ 安達 淳‡

† 東京大学大学院工学系研究科 ‡ 国立情報学研究所

〒 101-8430 東京都千代田区一ツ橋 2-1-2 国立情報学研究所

TEL: 03–4212–2681 E-mail: {tkana, akiko, takasu, adachi}@nii.ac.jp

本稿では、我々が提案している検索対象の著者キーワード リストを利用した問い合わせの翻訳手法と検索精度の向上手法の組み合わせが言語横断検索において有効であることを、NTCIR 1/2 を用いた評価実験により示す。翻訳手法は、検索対象のコーパスから多言語キーワード クラスタを自動抽出して問い合わせ表現の翻訳に用いることで、対訳辞書を用いる手法における辞書整備の問題を克服するものである。また、検索手法（関連性の重ね合わせモデル）は自然言語の意味曖昧性に着目し、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。本稿の実験ではその効果の言語独立性を示す。

**キーワード** 情報検索，ベクトル空間モデル，文書ベクトル修正, RS モデル，NTCIR, 言語横断検索

# Cross-Language Information Retrieval Using the Relevance-based Superimposition Model

Teruhito KANAZAWA† Akiko AIZAWA‡ Atsuhiro TAKASU‡ Jun ADACHI‡

† Graduate School of Engineering, University of Tokyo

‡ NII (National Institute of Informatics)

NII, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN

TEL: +81–3–4212–2681 E-mail: {tkana, akiko, takasu, adachi}@nii.ac.jp

We proposed a query translation method using the keyword lists given by the authors of documents and a document feature modification method based on the relevance of documents. In this paper, we show the effectiveness of our IR methods in cross-language information retrieval by the experiments with NTCIR 1/2 multilingual corpora. The translation method uses multilingual keyword clusters derived from the corpora as a dictionary. It is expected to overcome the problem of dictionary update. The Relevance-based Superimposition (RS) model modifies the document feature vectors using document clusters organized based on the relevance of documents, and it is expected to achieve higher precision of retrieval, independent of language.

**keywords** information retrieval, vector space model, document vector modification, RS model, NTCIR cross-language IR

## 1 はじめに

World Wide Web の普及によって世界中から発信された情報へのアクセス可能性が飛躍的に向上した。それに伴い情報検索、特に言語横断検索に対する要求も高まっている。本稿では我々が提案している翻訳手法と検索精度向上手法の、言語横断検索における有効性を検証する。

言語横断検索では問い合わ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、問い合わせの翻訳、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。以上の理由から問い合わ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、情報検索の問い合わせは極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。すなわち問い合わ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 [1, 2, 3]。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り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。そこで本稿では NTCIR コーパスが文献単位の対応関係を持っている対訳コーパスである点に着目し、著者キーワードから多言語キーワードクラスタを生成して、これを辞書に用いて問い合わせ表現の翻訳を行う手法を提案し、その評価を行う。

また、検索処理自体の精度の向上手法として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 [4, 5, 6]。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ある。本稿では NTCIR 1/2 テストセットを用いた言語横断検索（日本語問い合わ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。そして提案手法が言語横断検索において、言語あるいは検索対象データベースに依存した複雑なパラメータチューニングを必要とせずに一般的な tf·idf よりも高い精度を達成できることを示す。

本稿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 章で問い合わせ表現の翻訳手法について述べ、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察を 5 章で行い、最後にまとめを述べる。

## 2 RS モデル

検索対象の文献と問い合わせ表現は共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い合わせとは無関係な文書が検索されたり、同義多表記によって検索されるべき文書が検索できない場合がある。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一つであり、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) など問い合わせ表現を補正するもの、主成分分析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い合わせ表現に関連する語句を加えることで問い合わせの特徴ベクトルを拡張するものである。問い合わ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 [7]。

特徴空間を変化させるアプローチは Latent Semantic Index [8] などの手法に代表されるように、主成分分析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。こられらの手法の課題は主成分分析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、問い合わ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、問い合わ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。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。ここで言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、一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、例えば「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。提案手法では一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。

### 2.1 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．例えば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。

### 2.2 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．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．$\alpha$-関数族 [9] から派生する幾つかの関数の評価 [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，$d_{j,i}$ は文書 $\mathrm{d}_j$のベクトル$\boldsymbol{d}_j$の第 $i$ 要素である．

### 2.3 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．文章が属する全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毎に比較して，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．

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

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

ただし，$r_{1,i}, \cdots, r_{m,i}$ は文書 $\mathrm{d}_j$が属す文書群 $\boldsymbol{r}_1, \cdots, \boldsymbol{r}_m$の代表ベクトルの第 $i$ 要素である．

## 3 問い合わせ表現の翻訳

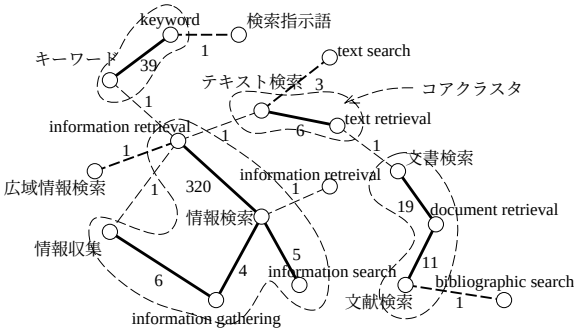
### 3.1 多言語キーワード クラスタの作成

筆者らはグラフ理論に基づいた多言語キーワード クラスタの生成手法を提案している [11, 12]。クラスタの生成は、コーパスからキーワードの対訳対を抽出することから始まる。本稿における実験で用いた NTCIR コーパスは各文献の著者が付与した和英の自由キーワードを含んでおり（図 1)、文献単位で和英の対応関係が記録されている。個々の著者キーワードの言語対応関係は必ずしも保証されないが、予備実験として無作為に選んだ 1000 対のうち 93%が意味的に対応していることが分かった [12]。そこでまず、コーパス中に含まれるキーワードをノード、対訳関係をリンクとみなして、コーパス全体をキーワードグラフによる表現に置き換える。図 2 にその一部を示す。この状態では、本来対応しないキーワード間を連結するノイズが存在し、巨大なクラスタが作成されてしまう。本手法ではこの点に着目し、連結するキーワード クラスタを分割するような対訳リンクの集合、すなわちグラフ理論における辺カットが対訳誤りの候補となることを仮定して誤り候補の検出を行う。

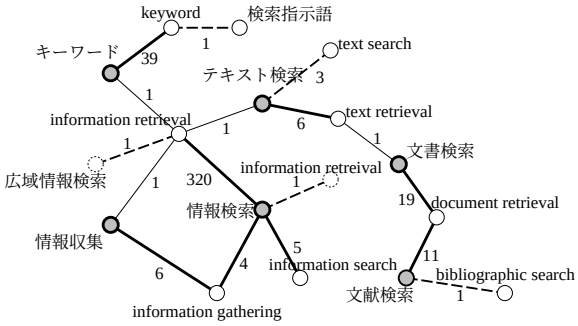
対訳グラフは 2 種類の誤りを含んでいると考えられる。図 2 における〈キーワード，*information*

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

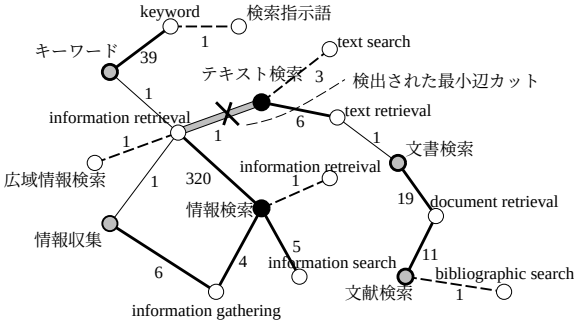
**図 1**　著者キーワードの例



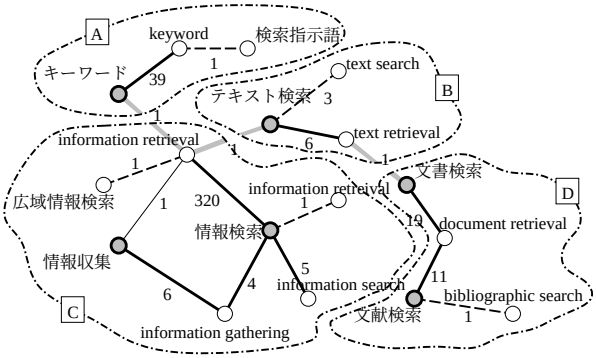

Step (1): 削除不能なリンクのチェック




Step (2): 重要語のチェック（図中では和用語のみ強調）




Step (3): 2 つの重要語の間の最小辺カットの検出




最終的な分割結果

**図 2**　クラスタ分割手順

*retrieval*⟩ のような完全な誤りと、⟨ テキスト検索, *information retrieval*⟩ のように対訳とはみなせないが関連性を持つ対である。前者は機械的に分割してよいリンクだが、後者は利用目的や状況にも依存しており、その判定は専門家でも難しい。例えば ⟨ テキスト検索, *information retrieval*⟩ は専門用語を定義する立場からは不適切であるとしても、情報検索の立場からは必ずしも誤りではない。このような点を考慮して提案手法では、コーパス内での出現頻度を手がかりに、以下の手順に従ってクラスタ分割の条件を設定する

### 3.1.1　削除不能な対訳リンクのマーク

次のいずれかに当てはまる対訳リンクは正しいと仮定して削除の対象から除外する。

1. 和英用語が同一表記。
2. コーパス中での出現頻度が $N_\alpha$ より大きい。
3. 和英いずれかのキーワードについて、その対訳リンク自身が唯一の対訳関係にある。

### 3.1.2　重要語のマーク

以下の条件を満足する語は重要語としてマークする。

1. 2 つ以上の対訳を持つ。
2. 頻度が $N_\beta$ よりも大きい対訳を少なくとも 1 つ持つ。

#### 3.1.3 最小辺カットの検出

最後に、重要語同士を分割する最小辺カットを検出し、以下を満足する場合に削除の対象とする。

- 最小辺カットを構成する対訳リンクの出現頻度の和が $N_\varepsilon$ 以下。

すなわち、最小辺カットの容量が $N_\varepsilon$ よりも大きい場合は、クラスタ内の重要語同士の結び付きが強いことから、これらの語は類義であると見なす。

以上をまとめると、クラスタの分割において、予め指定するパラメータは $N_\alpha, N_\beta, N_\varepsilon$ の 3 つである。これらのパラメータによって指定される分割条件に従って、全てのクラスタについてそれ以上行えなくなるまで再帰的に分割を繰り返す。$N_\alpha$ の値はコーパスの性質に応じてクラスタによらず定まると考えられる。一方、$N_\beta$ の値が小さいほど、あるいは $N_\varepsilon$ の値が大きいほどクラスタの分割が進むことが予想できる。本稿の実験では $N_\alpha = 3, N_\beta = 3, N_\varepsilon = 10$ に固定している。この時、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。

### 3.2 キーワードクラスタを用いた問い合わせの翻訳

生成された対訳キーワードクラスタを用いて、問い合わせ表現の翻訳を行う。この処理は以下の手順を繰り返すことで行われる。

1. 日本語の問い合わせ表現の先頭から、最長一致のキーワードを探す。
2. 最長一致のキーワードを含むクラスタが定まる。
3. クラスタ内で出現頻度が最大の英語キーワードを対訳とする。

## 4 評価実験

評価実験は NTCIR 2（予備版）の問い合わせ#101〜#149 を用いて、NTCIR の標準的な評価方法に従い、問い合わせごとの最大上位 1000 件における再現率と適合率を求めた [14]。以下、J-J タスクとは日本語問い合わせによる日本語文書の検索、E-E タスクとは英語問い合わせによる英語文書の検索、J-E タスクとは日本語問い合わせによる英語文書の検索をさす。

表 1 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。

**表 1** コーパスから抽出されたキーワードの数

| | 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 検索システム $R^2D^2$

$R^2D^2$(RetRieval system for Digital Documents) はベクトル空間モデル [15] に基づいて作成した文献検索のためのシステムである [6]。本稿における評価実験は $R^2D^2$ を用いて、一般的な tf·idf 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。

#### 4.1.1 形態素解析

日本語の形態素解析は chasen 1.51[16] 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 [13] 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。

#### 4.1.2 検索語の重み付け

$R^2D^2$ では検索語 $\{q_0, q_1, ..., q_m\}$ から成る問い合わせ $\boldsymbol{Q}$ 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 [17] 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、予備実験で性能のよかった、

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

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 $\Delta_i$ として、

$$c(i) \equiv \sum_{d_j\in\Delta_i}\sum_{t_k\in d_j} f_D(k) \quad (6)$$

$$\overline{c}(i) \equiv \sum_{d_j\notin\Delta_i}\sum_{t_k\in d_j} f_D(k) \quad (7)$$

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

を用いた。式 (8) では、問い合わせの話題に関連度の高い文書集合における情報量 $\log\frac{c(i)}{df_i}$と、補集合における情報量 $\log\frac{\overline{c}(i)}{N-df_i}$ 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$\sum_i = w(i,j,\boldsymbol{Q})$ とする。

### 4.2 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い合わ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 relevance feedback に基づく自動 QE[13] 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。言語横断検索では、翻訳した問い合わせに対して同様の操作を行った。NTCIR の問い合わせ#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 の他の問い合わせを用いて、パラメータと QE の性能の関係を調べる。

### 4.3 結果

表 2 に各手法の検索精度を示す。E-E タスクにおける 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。J-E タスクの baseline は 0.2401 で、QE、RS モデルの寄与は E-E タスクと同程度であった。J-J、E-E、J-E のいずれのタスクにおいても RS モデルと QE を組み合わせた場合の精度向上率 7〜9%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る。

表 3 は各手法の特徴を問い合わせ毎の平均適合率についての統計値で表したものである。baseline と RS モデルの平均適合率の差、すなわち ‘RS モデルの寄与’ の平均は baseline と QE 間の差、すなわち ‘QE の寄与’ の約 3 倍であるのに対し、RS モデルの寄与の偏差は QE の寄与の偏差の約半分である。これらの値は RS モデルの寄与が QE よりも正方向に偏っていることを意味しており、問い合わせによらず検索精度を向上させることを示すものである。一方、QE の寄与は NTCIR 1/TREC と NTCIR 2 とで大きく異なっており、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。また、表 2 の結果は RS モデルが問い合わせの言語に依存せず有効であることを示している。

表 4 は問い合わせの翻訳誤りを分類したものである。分類は提案手法による翻訳結果と NTCIR の E-E タスク用問い合わ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。全体の約 75%がこれらに含まれる。一方、翻訳誤りと分類される D〜G の個々の事例を調べると、「各種抗菌物質の MRSA に対する効果について」という問い合わせ中の「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が多いことが分かった。総合すると単一言語検索である E-E タスクに比べて言語横断検索の J-E タスクは平均適合率による比較で約 8 割の精

**表 2**　各手法の平均適合率

| 手法 | QE のパラメータチューニング NTCIR 1 J-J | TREC 3 | NTCIR 2 J-J | E-E | J-E | /E-E |
|---|---|---|---|---|---|---|
| baseline | .3059 | .2318 | .2841 | .2984 | .2401 | 0.80 |
| QE | .3270 (+7%) | .2578 (+11%) | .2886 (+2%) | .3044 (+2%) | .2441 (+2%) | 0.80 |
| RS | | | .3020 (+6%) | .3160 (+6%) | .2508 (+4%) | 0.79 |
| QE+RS | | | .3103 (+9%) | .3226 (+9%) | .2574 (+7%) | 0.80 |

**表 3**　各手法の統計量 (J-E タスクにおいて)

| 手法 | 精度への寄与の平均 | QE に対し | 精度への寄与の標準偏差 | QE に対し |
|---|---|---|---|---|
| QE | +0.0029 | 100% | 0.0606 | 100% |
| RS | +0.0107 | 369% | 0.0306 | 50% |
| QE+RS | +0.0176 | 607% | 0.0686 | 113% |

**表 4**　自動生成キーワードクラスタを用いた問い合わせの翻訳の誤り分類

| 分類 | 語数 | | 例 |
|---|---|---|---|
| A | 149 | (54%) | gravity, natural language processing |
| B | 33 | (12%) | genetic engineering techniques → genetic engineering |
| C | 26 | ( 9%) | heart disease → coronary artery disease |
| D | 8 | ( 3%) | tomography → imaging |
| E | 21 | ( 8%) | Historical materials on the Internet → historical materials, internet, electronics, database |
| F | 37 | (13%) | Distace education support system |
| G | 4 | ( 1%) | US → rice |
| 計 | 278 | | |

矢印の左は E-E タスクの問い合わせ表現 ($Q_m$)、右は提案手法による翻訳 ($Q_a$)。分類は、$Q_m$と $Q_a$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) 全く異なる意味の表現。

度であるという結果が得られ、RS モデルや QE を適用した場合でもほぼ同じ比率であった。

## 5　考察

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、本稿で述べ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挙げられる。本稿で述べた翻訳手法、RS モデルはそれぞれ対訳関係の抽出と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る。前者は検索対象のデータベースにキーワードが付与されていない場合でも他のコーパスに対訳キーワードがあればそれを用いることが可能である。一方、後者は検索対象となっている文書間の関係を解析する必要があるので、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ことになる。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。

## 6 おわりに

本稿は NTCIR テストセットを用いた実験結果により、RS モデルの言語非依存の効果と、グラフ理論に基づいて自動生成した多言語キーワードクラスタが問い合わせ表現の翻訳に有効であることを示した。

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%の精度向上を達成した。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 2 が示すように、二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 が適切に問い合わ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。

グラフ理論に基づいて生成した多言語キーワードクラスタを用いて問い合わせを翻訳した言語横断検索手法は、単一言語検索の 80%程度の検索精度を得た。これは他の言語横断検索手法に比肩するものである [18]。

## 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。また、「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。

## 参考文献

[1] A. Pirkola, "The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval," *SIGIR '98*, pp. 55–63, 1998.

[2] L. Ballesteros and W. Bruce Croft, "Resolving Ambiguity for Cross-Language Retrieval," *SIGIR '98*, pp. 64–71, 1998.

[3] S. Fujita, "Notes on the Limits of CLIR Effectiveness: NTCIR-2 Evaluation Experiments at Justsystem," *NTCIR Workshop 2 Proc.*, pp. 5–75–5–82, Tokyo, March 2001.

[4] T. Kanazawa, "$R^2D^2$ at NTCIR: Using the Relevance-based Superimposition Model," *NTCIR Workshop 1 Proc.*, pp. 83–88, Tokyo, Aug. 1999.

[5] T. Kanazawa, A. Takasu, and J. Adachi, "A Relevance-based Superimposition Model for Effective Information Retrieval," *IEICE Transactions*, 2000 (to appear).

[6] T. Kanazawa, A. Takasu, and J. Adachi, "$R^2D^2$ at NTCIR 2 Ad-hoc Task: Relevance-based Superimposition Model for IR," *NTCIR Workshop 2 Proc.*, pp. 5–98–5–104, Tokyo, March 2001.

[7] M. Mitra, A. Singhal, and C. Buckley, "Improving Automatic Query Expansion," *SIGIR '98*, pp. 206–214, 1998.

[8] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "indexing by latent semantic analysis," *J. American Society for Information Science*, Vol. 41, No.6, pp. 391–407, 1990.

[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), Vol. 98, No.58, 98-DBS-116(2), pp. 381–388, July 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, March 1999.

[11] A. Aizawa and K. Kageura, "An Approach to the Automatic Generation of Multilingual Keyword Clusters," *Proc. COMPTERM'98*, pp. 8–14, 1998.

[12] 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.

[13] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.

[14] NTCIR staff, "Notes on Evaluation for Japanese & English IR Tasks," *NTCIR Workshop 2 Proc.*, pp. 6–117–6–120, March 2001.

[15] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[16] "Japanese Morphological Analyzer 'ChaSen'*(in Japanese)* , `http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html`.

[17] 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), Vol. 2000, No.69, 2000-DBS-122, pp. 57–64, 岩手, July 2000.

[18] N. Kando, K. Kuriyama, M. Yoshioka, "Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop," *NTCIR Workshop 2 Proc.*, pp. 4–37–4–59, March 2001.