Webディレクトリを言語資源として利用した言語横断情報検索

木村文則[†] 前田亮[††] 宮崎純[†]
吉川正俊[†††] 植村俊亮[†]

インターネットの世界的な普及により，言語横断情報検索の重要性が増している．これまでに様々な手法が研究され，問合せの翻訳や訳語の曖昧性解消などにコーパスを利用する手法などにより一定の成果が得られている．しかしこのような手法では，コーパスの分野に対する依存が大きいため，コーパスが対象としていない分野に対しては検索精度が低くなる可能性がある．そこで本論文では，Web情報の言語横断情報検索において，たとえばYahooのような複数の言語版を持つWebディレクトリを利用する手法を提案する．事前に，カテゴリごとに属するWeb文書から特徴語を抽出し，これを比較することにより対応する異言語のカテゴリを決定する．検索において問合せが与えられると，問合せが適合する同言語のカテゴリに対応する異言語のカテゴリの特徴語を用いて問合せを翻訳することにより，訳語の曖昧性解消を図る．また，提案手法による検索の実験を行い，有効性の検証を行った．曖昧性解消を行わない対訳辞書による問合せ翻訳の場合よりも，提案手法の方が検索精度が向上することが明らかになった．

Cross-Language Information Retrieval Using Web Directories as a Linguistic Resource

Fuminori Kimura,[†] Akira Maeda,[††] Jun Miyazaki,[†]
Masatoshi Yoshikawa[†††] and Shunsuke Uemura[†]

With the popularity of the Internet, more and more languages are used for Web documents. Since the Web consists of documents in various domains or genres, the method for Cross-Language Information Retrieval (CLIR) of Web documents should be independent of a particular domain. In this paper, we propose a CLIR method which employs Web directories provided in multiple language versions (such as Yahoo). In the proposed method, feature terms are first extracted from Web documents for each category in the source and the target languages. Then, one or more corresponding categories in another language are determined beforehand by comparing similarities between categories across languages. Using these category pairs, we intend to resolve ambiguities of simple dictionary translation by feature term set of the categories to be used for disambiguation. In order to verify the effectiveness of our method, we conducted experiments of the proposed retrieval method using English and Japanese versions of Yahoo. This experiment proved that the proposed method is more effective for CLIR than simple dictionary translation without disambiguation.

1. はじめに

世界的なインターネットの発展にともない，外国語文書を電子的に入手することが容易となった．しかし従来のWeb検索エンジンは，問合せと同一言語の文書群が検索対象であるため，外国語文書に対する検索は効率的とはいい難い．

また，利用者の検索要求によっては，利用者の母国語以外の言語で記述された情報の方が豊富である場合も考えられ，これらを検索したいというニーズは少なくないと思われる．従来の単言語検索システムにおいてこのような要求を満たすには，利用者自身が辞書などを用いて問合せを翻訳する必要がある．この作業は利用者に負担を強いるだけでなく，不慣れなあるいはまったく読み書きができない言語に翻訳する場合は，適切な訳語の選択を誤る可能性が高い．

このような要求から，ある言語で書かれた文書群を

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†† 立命館大学情報理工学部メディア情報学科
Department of Media Technology, College of Information Science and Engineering, Ritsumeikan University
††† 名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University

別の言語による問合せで検索することを可能とする言語横断情報検索（Cross-Language Information Retrieval: CLIR）に関する研究が近年さかんになってきている．言語横断情報検索に関する従来の研究では，問合せの翻訳や訳語の曖昧性解消などにコーパスを利用する手法などが提案され，検索精度の向上において一定の成果が得られている．しかしコーパスを利用した手法では，学習に用いるコーパスのドメインに対する依存が大きいため，それ以外のドメインに対しては検索精度が低くなる可能性がある．Web 文書の言語横断検索では文書内容の分野は広範囲にわたっているため，ドメイン依存の問題を改善しなければならない．

そこで我々はこれまでに，Web 情報の言語横断情報検索において，たとえば Yahoo のような複数の言語で類似の構造を持つ Web ディレクトリを利用する手法を提案している[1)~3)]．まず，カテゴリごとに属する Web 文書から特徴語を抽出し，これを比較することなどにより対応する異言語のカテゴリを決定する．検索を行うときは，まず問合せと適合するカテゴリを同一言語間で選択し，次にそのカテゴリに対応する異言語のカテゴリを選択する．この選択された異言語のカテゴリの特徴語を利用することにより，問合せの翻訳における曖昧性解消を行う．こうして得られた問合せの訳語を用いることで，目的の文書群に対して検索を行う．このような方法により訳語の曖昧性解消を行い，言語横断情報検索の性能の向上を図る．

本論文の 2 章で関連研究について触れる．3 章において，提案手法について説明する．4 章では，提案手法による検索の実験を行う．最後に 5 章において結論を述べる．

2. 関 連 研 究

言語横断情報検索に用いられる手法は大きく分けて，検索対象の文書群を翻訳する方式，問合せを翻訳する方式，言語に依存しない中間言語を用いる方式の 3 つがある．

検索対象の文書群を翻訳する方式は，既存の機械翻訳システムを用いることができ，文脈を考慮できることにより訳語の曖昧性も低くなることから，一般に問合せを翻訳する方式より高い検索精度が得られるとされている[4)]．しかしながら，大規模な文書群をすべてあらかじめ翻訳しておくことは現実的ではなく，対応言語の拡張も困難であるため，Web のように多言語が混在し，かつ大規模で更新が頻繁な文書群の検索には不向きである．

問合せを翻訳する方式においては，特に Web 検索エンジンの一般的な利用者が投入する問合せは平均 2 単語程度と短く，単語の羅列である場合が多いため[5)]，訳語の曖昧性の解消が問題になる．しかしながら，この方式は，翻訳された問合せを既存の単言語検索エンジンでそのまま用いることができるという利点がある．この方式では，まず対訳辞書を用いて問合せを翻訳し，これに対して訳語の曖昧性を解消する．本研究で用いる手法もこの範疇である．

訳語曖昧性解消にコーパスを用いる手法では，検索要求とコーパス間のドメインの相違による検索精度への影響が指摘されている．Hull[6)]および奥村ら[7)]は，並列コーパスや類似コーパスを用いる手法において，検索要求とコーパス間のドメインの相違が検索精度に悪影響を及ぼす可能性があることを指摘している．また Lin ら[8)]は，単言語コーパスとしてドメインや規模の異なる 3 つのコーパスを用いて比較実験を行った結果，有用な共起情報を得るには大規模でドメインの一致したコーパスが必要であると結論付けている．

本研究で対象とする Web 検索では，多様な分野の検索要求に対応することが要求される．しかし，そのそれぞれのドメインについて，対応するコーパスをあらかじめ用意することは現実的ではない．本研究では，Yahoo などの複数の言語版が用意されている Web ディレクトリに登録されている文書群をコーパスとして用い，これを訳語の曖昧性解消に用いることで言語横断情報検索を行う手法を提案する．Web ディレクトリには多様な分野の Web 文書が登録されているため，Web ディレクトリをコーパスとして用いることは，ほとんどのドメインに対応したコーパスを利用することであるといえる．これにより，問合せが対象とする分野に依存しない言語横断情報検索システムの実現が可能となる．

3. 提案する手法

図 1 は提案手法のシステムの概要を表している．本システムは，問合せおよび検索対象のそれぞれと同じ言語版の Web ディレクトリ，それぞれの言語の特徴語データベース，対訳辞書，検索対象となる文書群から構成される．図 1 において点線で囲まれている部分は，問合せの翻訳処理の構成を表している．

本システムは，Web ディレクトリの各カテゴリから特徴語を抽出してそれを特徴語データベースに事前に格納しておく前処理と，与えられた問合せを翻訳して検索を行う検索処理の 2 つの処理に分けられる．

3.1 前 処 理

図 2 は前処理の流れを示したものである．前処理と

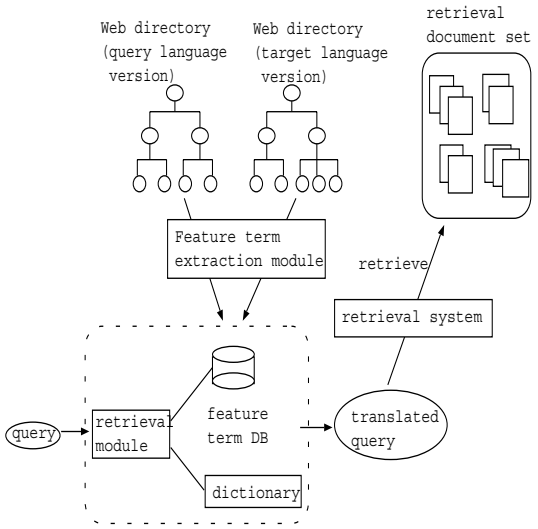

図 1 提案するシステムの概要
Fig. 1 Outline of the proposed system.

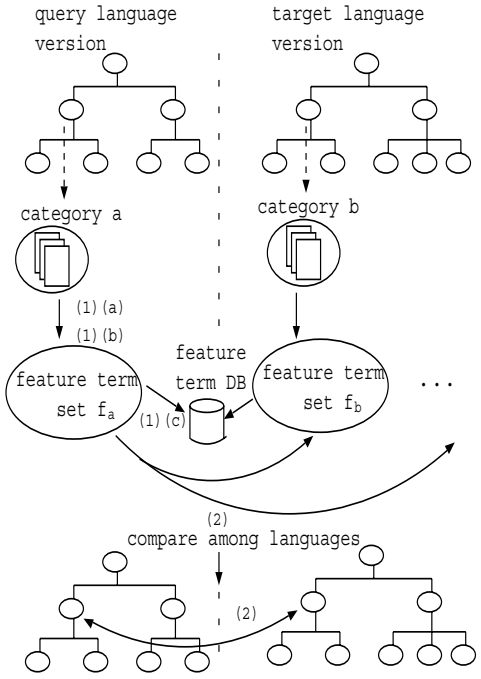

図 2 前処理における処理の流れ
Fig. 2 Flow of preprocessing.

して事前にそれぞれのカテゴリにおいて，特徴語の抽出と異言語のカテゴリとの対応付けを行う．前処理の手順を以下に示す．

(1) 特徴語の抽出
各言語版の Web ディレクトリのすべてのカテゴリに対して
(a) そのカテゴリに属する Web 文書から単語を抽出し，重み付けを行う．
(b) 重みの大きい上位 n 語の単語をそのカテゴリの特徴語として抽出する．
(c) 抽出された特徴語を特徴語データベースに格納する．

(2) 言語間でのカテゴリの対応付け
すべてのカテゴリに対して対応する異言語のカテゴリを推定し，対応付ける．

たとえば図 2 の query language version のカテゴリ a に対する対応付けでは，まずカテゴリ a に属する文書群から単語を抽出し，それらのカテゴリ a における重みを計算する (1)-(a)．次に，抽出された単語のうちから重みの大きいものから n 語を特徴語として抽出し，特徴語集合 f_a を得る (1)-(b)．こうして得られた特徴語集合 f_a を特徴語データベースに格納する (1)-(c)．得られた特徴語集合 f_a に最も類似していると思われる target language version のカテゴリを探し，カテゴリ a とそのカテゴリに対して対応付けを行う (2)．なお，対応付けの方法はどのような方法によって行ってもよい．たとえば，カテゴリの特徴語を比較することにより対応付けを行うことが考えられる．また，人手により直接対応付けを行うことも考えられる．こうして得られたカテゴリの対応は，文書の検索を行うときに利用する．

3.1.1 特徴語の抽出

各カテゴリは特徴語集合によりその特徴を表現される．特徴語集合は，そのカテゴリの特徴を表現していると思われる単語の集合である．特徴語を抽出するために，まず各カテゴリに属する Web 文書から単語を抽出する．次に，抽出された単語をカテゴリごとに集計し，その単語がカテゴリの内容を表現する程度を表す重みを計算する．抽出された単語のうち，重みが大きいものをそのカテゴリの特徴語として抽出する．

Web 文書から抽出された単語の重みは，TF·ICF（term frequency · inverse category frequency）により計算する．これは，一般によく知られた単語の重み付けの手法の 1 つである TF·IDF（term frequency · inverse document frequency）を発展させたものである．TF·IDF は単語の出現頻度（TF）と文書頻度の逆数との積により求められる．TF は単語の網羅性を表し，IDF は単語の特定性を表しており，これらの積である TF·IDF は網羅性と特定性がともに高い単語の重みが大きくなるようになっている．TF·IDF は，次の式で求められる．

表 1 英語版 Yahoo におけるカテゴリ “Government” の特徴語（上位 10 語）
Table 1 Top 10 feature terms of the category “Government” in English version of Yahoo.

特徴語	重み
law	0.001908
font	0.001560
court	0.001381
var	0.001233
information	0.001142
document	0.001136
war	0.001124
px	0.001014
time	0.000958
government	0.000938

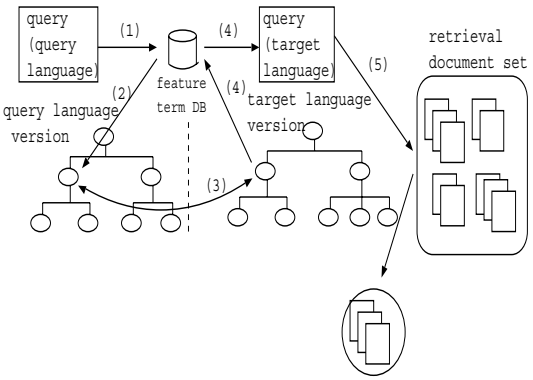

図 3 検索の流れ
Fig. 3 Flow of retrieval.

$$tf \cdot idf(t_i, d) = \frac{f(t_i)}{N_d} \cdot \log\left(\frac{N}{n_d} + 1\right)$$

ここで，$f(t_i)$ は特徴語 t_i の出現頻度，N_d は文書 d の全単語数，N は全文書数，n_d は特徴語 t_i が出現する文書数を表す．

TF·IDF では文書を単位として重みを計算するが，文書のかわりにカテゴリを単位として重みを計算するのが TF·ICF である．TF·ICF により重みを計算することで，文書単位で計算する TF·IDF より，カテゴリの内容をより反映した重み付けを行うことができる[9]．**表 1** は，英語版 Yahoo におけるトップレベルカテゴリ “Government” の特徴語の上位 10 語を表している．表 1 の単語は他の 12 カテゴリの特徴語にも含まれているが，カテゴリごとで重みが異なっているため，同じ特徴語でもカテゴリによって重要度も異なっている．

3.2 検索処理

本システムにおける検索の流れを**図 3** に示す．まず，問合せの適合カテゴリを選択し，続いて適合カテゴリに対応付けられている異言語のカテゴリを選択し，そのカテゴリの特徴語集合を利用して問合せの翻訳を行い，最後に翻訳された問合せを用いて文書群に対して検索が行われる．検索における処理の手順は次のようになる．

(1) 問合せと同じ言語版のすべてのカテゴリに対して問合せとカテゴリの特徴語集合との適合度を求める．
(2) 最も適合度の高いカテゴリを問合せの適合カテゴリと決定する．
(3) 検索対象の言語版のカテゴリから，適合カテゴリに対応付けられているカテゴリを選択する．
(4) 選択された対応カテゴリの特徴語集合を利用して問合せを翻訳する．
(5) 翻訳された問合せにより，検索対象の文書群を検索する．

3.2.1 問合せの適合カテゴリの選択

本システムにおける問合せは文章ではなく，数語の単語から構成されていることを前提としている．ここで，$t_1, t_2, \cdots, t_n$ の単語から構成される問合せ q に対する問合せベクトル $\vec{q}$ を次のように定義する．

$$\vec{q} = (q_1, q_2, \ldots, q_n)$$

なお，q_k は問合せの k 番目の単語 t_k に対応しており，その値は 1 である．

与えられた問合せについて，まず同言語間において，問合せと各カテゴリとの適合度を計算し（図 3 (1)），そのうちから最も適合度が高くなるカテゴリを，問合せが適合する同言語のカテゴリと決定する（図 3 (2)）．問合せとカテゴリの適合度は，問合せベクトルとカテゴリの特徴語集合のベクトルの内積にこの 2 ベクトルのコサイン距離を掛けることにより計算する．ここで，カテゴリ c の特徴語集合のベクトル $\vec{c}$ を，次のように定義する．

$$\vec{c} = (w_1, w_2, \ldots, w_n)$$

なお，w_k は，単語 t_k のカテゴリ c における特徴語の重みを表す．

問合せとカテゴリの適合度 $rel(q, c)$ は次のように求められる．

$$rel(q, c) = (\vec{q} \cdot \vec{c}) \frac{\vec{q} \cdot \vec{c}}{|\vec{q}| \cdot |\vec{c}|}$$

2 つのベクトルの内積のみから適合度を求めると，次のような場合に問題が生じる．いくつかある問合せ語のうちの 1 つだけしか特徴語集合に存在していないが，その重みが大きいという場合である．この場合，それ以外の問合せ語がその特徴語集合に存在していなくても，その存在している問合せ語の重みが大きいため，適合度の値が大きくなることがある．特徴語集合

に存在している問合せ語の数が多く，かつそれらの重みが高い，という 2 つの要求の両方を満たしている度合いが高いほど，$rel(q, c)$ の値も高くなるのが理想である．しかし，1 つの問合せ語の重みのみが大きい場合では，上記の要求の前者を満たしていない．そこでベクトルのコサイン距離を掛けることで，特徴語集合に存在している問合せ語の数を考慮する[10]．

こうして求めた適合度が最も高いカテゴリを，問合せに対する適合カテゴリとする．本論文では適合カテゴリを 1 つだけ選択したが，適合度が閾値以上となるカテゴリを適合カテゴリとする方法も考えられる．このとき適合度が閾値以上となるカテゴリが複数ある場合は，これらをすべて適合カテゴリとして選択する．

次に，前処理で得られた対応付けからそのカテゴリに対応する異言語のカテゴリが決まる（図 3 (3)）．こうして得られた異言語のカテゴリの特徴語集合および対訳辞書を利用して，3.2.2 項で述べる方法により問合せを翻訳する（図 3 (4)）．こうして得られた問合せを用いて検索対象文書に対して検索を行う．以上の処理を経て得られた文書群が検索結果となる（図 3 (5)）．

3.2.2 問合せの翻訳

問合せの翻訳の流れを**図 4** に示す．まず，問合せ中の各単語 q に対する対訳辞書のすべての訳語 t_1，t_2，t_3，$\cdots$ を，訳語の候補として抽出する．抽出されたすべての訳語候補について，適合カテゴリに対応付けられている異言語のカテゴリ（以下，“対応カテゴリ”）b の特徴語に含まれているかを調べる．適合カテゴリの決定およびその対応カテゴリの決定方法については，3.2.1 項において述べた方法で行う．含まれていた訳語のうち，特徴語の重みが最も大きい訳語を，その問合せ語の訳語と決定する．このとき，対応カテゴリの特徴語集合の中にいずれの訳語候補も存在しない場合，その問合せ語は使用しない．しかし，たとえば，日本語で書かれた Web 文書中において英単語が使われるといったことも頻繁にあるため，翻訳を行わないほうがよい場合もある．そこで，いずれの訳語候補も比較している対応カテゴリの特徴語に含まれていない場合，翻訳する前の問合せの単語そのものが，比較している対応カテゴリの特徴語に含まれているかを調べる（図 4 点線）．もし含まれていれば，翻訳前の単語そのものをこの問合せ語の訳語と見なす．

たとえば，英語のカテゴリ “Computers and Internet” が問合せの適合カテゴリであるときに英語の “system” という単語の訳語を決定する場合を考える．“system” の訳語の候補として，“宇宙”，“方法”，“組織”，“器官”，“システム”，“系統”，$\cdots$ などが得られる．この訳語の候補のすべてに対して，適合カテゴリの対応カテゴリである日本語のカテゴリ “コンピュータとインターネット” の特徴語集合に存在するどうかを調べる．そのうち重みが最も高いもの，今回は “システム” を，英単語 “system” の訳語と決定する．もし，“system” のいずれの訳語候補も対応カテゴリの特徴語集合に存在しない場合は，“system” という単語そのものが対応カテゴリの特徴語集合に存在するか調べ，存在していれば “system” という単語そのものを訳語と見なす．

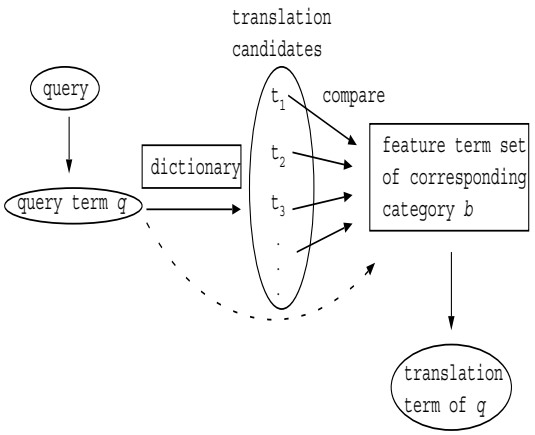

図 4 問合せの訳語の決定
Fig. 4 Translation of a query.

3.2.3 文書の検索

3.2.2 項で述べた手法により翻訳された問合せを用いて検索対象文書群に対して検索を行う．検索対象文書群は，必ずしも Web ディレクトリに登録されている文書でなくてもよい．検索システムは既存のシステムを使用することができる．こうして得られた文書群が，問合せに対する検索結果となる．

4. 評価実験

提案手法の有効性を検証するために，日本語の問合せから英語の文書群に対して検索する実験を行った．今回の実験の目的は，Web ディレクトリを利用した本手法が言語横断情報検索に対して検索精度の向上が得られるか，について検証することである．今回の実験では，提案手法だけではなく，比較対象として訳語の曖昧性解消を行わない場合についても，同様の実験を行った．比較対象における問合せの翻訳は，対訳辞書から得られる問合せ語の訳語をすべてを翻訳された問合せ語とすることで行う．ただし，複数の単語からなる訳語は，翻訳された問合せから省いている．問合せ翻訳後の処理については，提案手法と同様の方法で

```
<TOPIC>
<NUM>001</NUM>
<SLANG>CH</SLANG>
<TLANG>JA</TLANG>
<TITLE>展覧会「漢代の芸術と文化」</TITLE>
<DESC>
故宮博物館で行われた「漢代の芸術と文化」という展覧会についての
情報を探す．</DESC>
<NARR>
故宮博物館は，中国コレクションがすぐれていることでよく知られて
いる．漢代のコレクションは，紀元前 206 年から西暦 220 年までの
中国の繁栄期を表すものである．芸術的な展示物における文化的・歴
史的遺物の種類，展覧会の日程，故宮博物館がどのように展覧会の準
備をしたか，展覧会の協賛パートナー，展覧会に対する市民の反応な
どに焦点を当てている文書を関連文書と見なす．展示されていない漢
代の芸術や文化や他の展覧会の紹介は，不適合とする．
</NARR>
<CONC>
漢代，漢代の芸術と文化，展覧会，故宮博物館，歴史
</CONC>
</TOPIC>
```

図 5 NTCIR3 言語横断検索タスク 日本語検索課題（抜粋）
Fig. 5 Extract of a Japanese query in the NTCIR3 CLIR task.

検索を行う．

本実験では，国立情報学研究所が作成した，第 3 回 NTCIR ワークショップの言語横断検索タスクで用いられた文書群と検索課題（以下，NTCIR3 テストコレクション）を使用した．このテストコレクションのうち，1998〜1999 年に台湾で発行された英字新聞各種からなる EIRB010，および同年に日本で発行された英字新聞である “毎日デイリー 1998〜1999” の 2 つを検索対象として用いた．また，このテストコレクションの日本語の検索課題を本実験における問合せとして用いた．NTCIR3 の日本語検索課題は 50 の問合せが用意されており，このすべての問合せを用いて実験を行った．**図 5** は，NTCIR3 言語横断検索タスクの日本語検索課題を抜粋したものである．

また，訳語の曖昧性解消のために用いる Web ディレクトリとして，Yahoo の英語版と日本語版を用いた．本実験では，英語のトップレベルカテゴリ “Regional”，および日本語のトップレベルカテゴリ “地域情報” 以下のカテゴリを除いたすべてのカテゴリから Web 文書を収集し，曖昧性解消に用いた．英語のトップレベルカテゴリ “Regional”，および日本語のトップレベルカテゴリ “地域情報” 以下のカテゴリを除いたのは，これらのカテゴリには世界各地の地域に関する文書が属しているため，英語および日本語の翻訳に用いるのには適さないからである．英語版ではカテゴリ数は 84,835 カテゴリ，文書数は 800,000 文書，日本語版ではカテゴリ数は 3,175 カテゴリ，文書数は 34,443 文書であった．今回の実験では，下位のカテゴリを上位のカテゴリに統合し，最終的には各言語版のトップページに登録されている 13 のカテゴリに統合した．カテゴリの統合を行った理由は，カテゴリによっては属している Web 文書が少なく十分な統計情報が得られない場合もあるためである．統合後のカテゴリから抽出された単語数は，英語版では 1 カテゴリあたり 322,672 語であった．

Web 文書から単語を抽出する際に，英語版では単語の活用形を原形にしたのち，ストップワードを取り除いた．ストップワードのリストは “Information Retrieval: Data Structures and Algorithms” の chapter 7 [11] に掲載されているものを用いた．日本語では，英語のように単語の区切りが明確でないため，“茶筌”☆などの形態素解析ツールを用いる必要がある．本実験では “茶筌” を用いて単語に分割した後，名詞，動詞，形容詞，未知語を抽出した．また，問合せの翻訳のための対訳辞書には，“EDR 電子化辞書” の “日英対訳辞書”☆☆を用いた．単純に対訳辞書で翻訳した場合，1 単語に対して平均で 5.17 語の訳語候補が得られた．カテゴリの特徴語の抽出において，各カテゴリの特徴語数は 10,000 語とした．特徴語数の決定の詳細は 4.1 節において述べる．また，言語間におけるカテゴリの対応付けは人手により行った．Yahoo では，各言語版のトップページに登録されている 13 のカテゴリの構成は，いずれの言語でも同じであるので，対応が明らかであったためである．

前処理が済んだのち，NTCIR3 テストコレクションを用いて検索の実験を行った．問合せは，NTCIR3 言語横断検索タスクの日本語検索課題から “TITLE” フィールドを抽出したもの，および “DESC” フィールドを抽出したものの 2 通りを用いた．一般に，Web 検索エンジンの一般的な利用者が投入する問合せは平均 2 単語程度であるといわれている [5]．そのため，NTCIR3 テストコレクションの検索課題のフィールドのうちから比較的単語数の少ない “TITLE” および “DESC” フィールドを今回の実験で用いた．それぞれの問合せに対して “茶筌” により形態素解析を行い，名詞，動詞（サ変動詞の一部を除く），形容詞，未知語のみを抽出し，これらを問合せの単語として用いた．

問合せを翻訳したあとで行う検索処理においては，“SMART”☆☆☆を用いて検索を行った．検索対象文書群の索引は augmented tfidf により重み付けを行った．

☆ http://chasen.aist-nara.ac.jp/hiki/chasen/
☆☆ http://www2.crl.go.jp/kk/e416/EDR/j_index.html
☆☆☆ ftp://ftp.cs.cornell.edu/pub/smart/

4.1 問合せの翻訳

提案手法ではまず，問合せと同言語である日本語のカテゴリから適合カテゴリを推定する．このとき問合せと比較するカテゴリの特徴語集合は，各カテゴリにつき 1,000 語使用した．次に，推定された適合カテゴリに対応付けられている英語のカテゴリを選択する．そして，対訳辞書を用いて問合せ単語の訳語候補をすべて抽出する．最後に選択した英語の対応カテゴリの特徴語と問合せの訳語候補とを 3.2.2 項で述べた方法で比較することにより，訳語を決定する．

このとき，各カテゴリの特徴語数を事前に決定する必要がある．本システムにおいて，適切な特徴語数を調べるために，特徴語数の変化に対して訳語獲得率がどのように変化するかについて実験を行った．訳語獲得率とは，全問合せ語数に対して，何らかの訳語が得られた問合せ語数の割合である．今回の実験では，英語の対応カテゴリの特徴語数は，1,000 語から 10,000 語の間で，1,000 語ごとに実験を行った．

図 6 は実験の結果を示している．“TI”，“DE” はそれぞれ “TITLE” および “DESC” フィールドから得られた問合せを用いた場合の結果を示している．カテゴリの特徴語数が 1,000 語から 4,000 語の間では，訳語獲得率の上昇の度合いが大きい．さらに特徴語数を増加させていくと徐々に訳語獲得率の上昇の度合いは小さくなってゆき，9,000 語から 10,000 語ではほぼ変化はなくなった．カテゴリの特徴語数が 1,000 語の場合，TITLE では 44.4%，DESC では 57.0%と，全問合せ単語数の半分程度しか訳語が得られていない．特徴語数を 10,000 語にした場合，それぞれ 77.0%，83.0%の獲得率であり，特徴語数が 1,000 語の場合に比べて 30%程度獲得率が上昇している．

問合せの訳語が得られない原因として，対訳辞書に訳語候補が存在しない場合とカテゴリの特徴語にすべての訳語候補が存在しない場合の 2 点があげられる．このうち前者の原因は，カテゴリの特徴語数に関係なく，訳語の獲得率の低下を引き起こす．この実験で用いた問合せ語に対して，何らかの訳語が対訳辞書に存在している割合は，“TITLE” フィールドでは 88.8%，“DESC” フィールドでは 92.8%であった．ともに，特徴語数が 10,000 語の場合よりも約 10%高くなった．この差は，上記の原因の後者から生じていると考えられる．この後者の原因は，カテゴリの特徴語数がそのまま獲得率に影響する．よって，カテゴリの特徴語数を 10,000 より多くすることにより訳語の獲得率が上昇する可能性はある．しかし，特徴語数をあまり多くすると，そのカテゴリの特徴をあまり表していない単語が含まれてしまい，問合せに対して適切でない訳語を選択してしまうという問題が生じる可能性が考えられる．特徴語数が 10,000 語で訳語獲得率の上昇が小さいことも考慮すると，さらに特徴語数を増やすのはあまり有効ではないといえる．

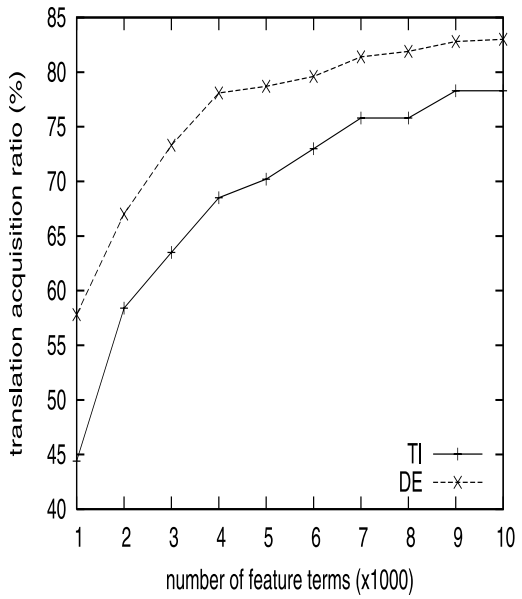

図 6 カテゴリの特徴語数と得られた訳語獲得率との関係
Fig. 6 Relation between aquired translation ratio and the number of feature terms in each category.

4.2 検　　索

翻訳された問合せにより，NTCIR3 言語横断検索タスクの英語文書群に対して検索を行った．本実験では，提案手法における問合せの翻訳に用いる各カテゴリの特徴語数は 10,000 語とした．その結果を**図 7** および**表 2** に示す．なお，文書が適合しているかどうかの判定は，NTCIR テストコレクションの Relax 正解集合を用いた．NTCIR3 テストコレクションには，“Rigid” と “Relax” の 2 種類の正解集合が用意されている．Rigid 正解集合では文書が適合しているかどうかの基準はやや厳格であり，それに比べて Relax 正解集合ではその基準をやや緩和している．

図 7 は検索結果の評価を適合率・再現率グラフで表したものである．この図における “TI-our”，“DE-our” はそれぞれ，“TITLE” フィールドを抽出した問合せ（以下，問合せ TI），および “DESC” フィールドを抽出した問合せ（以下，問合せ DE）を用いて提案手法により検索を行った結果である．“TI-base”，“DE-base” はそれぞれ，問合せ TI および問合せ DE を，対訳辞書のみによる翻訳を行った場合（以下，ベースライン）の検索結果である．適合率とは，検索結果として得ら

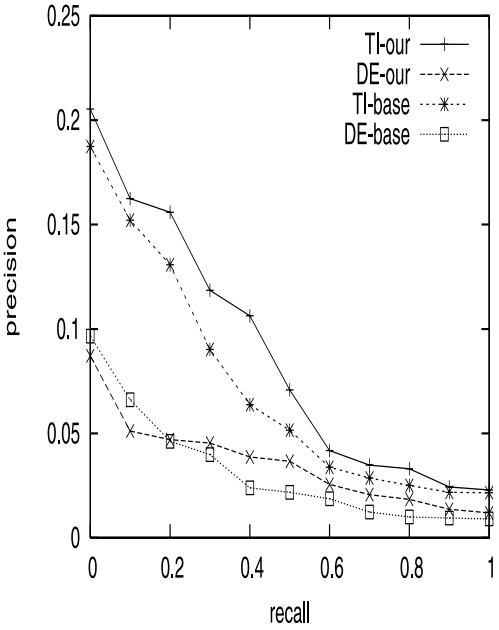

図 7 検索結果の適合率・再現率グラフ
Fig. 7 Precision-recall graph of retrieval results.

表 2 検索結果の平均適合率
Table 2 Average precision of retrieval results.

	TITLE	DESC
提案手法	0.0851	0.0311
ベースライン	0.0677	0.0254

れた文書のうち，実際に問合せに適合している文書の割合のことである．再現率とは，文書群全体に含まれている適合文書のうち，検索結果に含まれている割合のことである．また，表 2 は，上記 4 種類の問合せでの検索結果の 11 点平均適合率を表している．11 点平均適合率とは，再現率が $0.0, 0.1, 0.2, \cdots, 0.9, 1.0$ となる 11 点における適合率の平均値であり，その検索システムの検索精度を評価する指標の 1 つである．

4.3 考　　察

提案手法の総合的な検索性能の比較のため，4.2 節で得られた実験結果に対して考察を行う．表 2 から分かるように，問合せ TI においては提案手法がベースラインより 1.74 ポイント，問合せ DE では 0.57 ポイント平均適合率が上昇しており，問合せ TI，問合せ DE のいずれにおいても，本手法の有効性が認められる．図 7 においても，ほとんどの再現率の区間で，本手法のほうが良い結果が得られた．ただし，問合せ DE においては再現率が 0.00〜0.20 の区間においてのみ，本手法が下回る結果となった．この区間において本手法が下回った原因は，問合せの翻訳において訳語が一部得られなかったことにあると考えられる．一方ベースラインの場合，訳語候補はすべて用いるため，重要な問合せ語が含まれる確率も提案手法よりも高くなる．よって，再現率が低い段階では，重要な問合せ語が含まれていることの効果が強く現れたと考えられる．しかし，再現率が高くなってくると，重要な問合せ語が含まれることよりもむしろ，不要な問合せ語の影響が大きくなってくると考えられる．そのため，不要な語をできるだけ排し重要な語のみを訳語として用いる提案手法が，再現率が高くなる区間では良い結果が得られた．

問合せ TI と問合せ DE で比較すると，提案手法とベースラインのいずれにおいても，問合せ TI のほうが良い結果が得られた．“TITLE” フィールドは，重要な単語の羅列であるといってよく，そこに現れる単語のほとんどが重要な問合せ語であるといえる．しかし “DESC” フィールドは文章として完結した形となっているため，問合せには不要な語も含まれる．たとえば “〜に関する記事を探したい”，“〜について記述された文書を検索する” といった記述が多くあり，このような記述から抽出された単語が不要な問合せ語となった．このような不要な問合せ語が，検索精度の低下の原因となったと考えられる．このことからも，問合せから不要な単語を排除することが重要であるといえる．

NTCIR3 の CLIR タスクに参加した手法の多くは，0.2 から 0.3 程度の 11 点平均適合率が得られている[12]．これと比較すると，今回の実験結果の 11 点平均適合率はかなり低い．その原因は，固有名詞に対する対策をほとんどとらなかったことにある．固有名詞は適合文書を特定する重要な手がかりとなる．しかし固有名詞は対訳辞書に載っていないことが多いため，適切な訳語を得るのは困難である．そのため，今回の実験では多くの固有名詞の訳語が得ることができず，適合率が低くなったと考えられる．

そこで，提案手法により問合せを翻訳するときに，固有名詞のみ人手により翻訳した場合についても，同様の検索の実験を行った．**図 8** は，この実験結果を適合率・再現率グラフで表したものである．問合せ TI に対して，提案手法で翻訳した場合が “TI-our”，固有名詞を人手により翻訳した場合が “TI-our-proper” である．また，“TI-base-proper” は，ベースラインにおいて固有名詞を人手により翻訳した場合の検索結果である．固有名詞を人手により翻訳することにより，適合率が格段に向上している．また，**表 3** は，この実験における 11 点平均適合率を表している．固有名詞を人手により翻訳した場合の 11 点平均適合率は 0.2291 となった．この結果は，他の言語横断情報検索

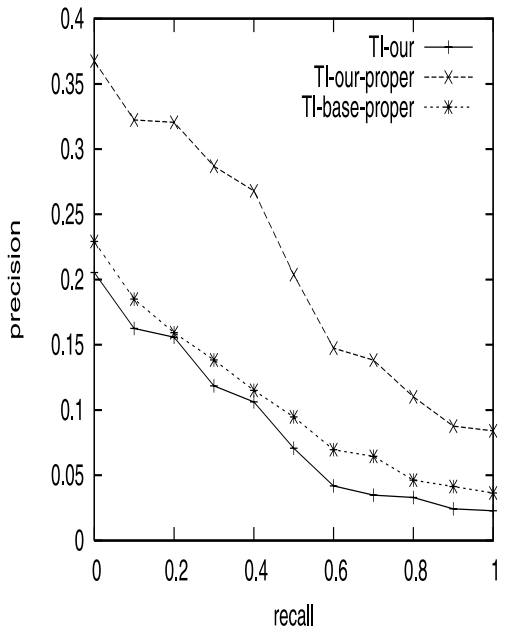

図 8 固有名詞を人手で翻訳した場合の適合率・再現率グラフ
Fig. 8 Precision-recall graph of retrieval results in case of manually translated proper nouns.

表 3 固有名詞を人手により翻訳した場合の検索結果の平均適合率
Table 3 Average precision of retrieval results in case of manually translated proper nouns.

TI-our	TI-our-proper	TI-base-proper
0.0851	0.2291	0.1022

の手法と比較しても遜色ない値であるといえる．また，“TI-base-proper” における 11 点平均適合率は 0.1022 であり，提案手法により検索精度が向することが確認された．

5. おわりに

本論文では，Yahoo に代表されるような，複数言語版が存在する Web ディレクトリを，言語横断情報検索における訳語の曖昧性解消と検索精度の向上に用いる手法を提案した．また，本手法の有効性を検証するために，NTCIR3 テストコレクションを用いて検索の実験を行い，本手法が言語横断情報検索の曖昧性解消において有効であることを示した．

本手法は，Yahoo などの複数の言語版が用意されている Web ディレクトリに登録されている文書群をコーパスとして用いることにより，分野に対する依存性が生じることはない．この特徴は Web 文書のように様々な分野が対象となる検索において有効であると思われる．また，本手法で用意すべき言語資源は対訳辞書のみであり，それ以外に特別に必要となる言語資源はない．さらに，Web ディレクトリには多数の言語版があるが（たとえば Yahoo は 2003 年 12 月の時点で 23 カ国版が存在），対訳辞書さえあればそれらの言語のすべての組合せに対して本手法は適用できるため，対応言語の拡大が容易である．

本研究の今後の課題として，カテゴリの統合方法の検討，対応言語の他の言語への拡大などがあげられる．

謝辞 本研究の一部は，日本学術振興会の科研費基盤研究（A）（2）「セマンティック Web のための多言語処理」（課題番号：15200010）の支援によるものである．ここに記して謝意を表す．

参考文献

1) 木村文則，前田 亮，吉川正俊，植村俊亮：ディレクトリ型検索エンジンのカテゴリ間対応付けによる言語横断検索，第 13 回データ工学ワークショップ（DEWS2002）(2002). http://www.ieice.org/iss/de/DEWS/proc/2002/papers/C4-4.pdf
2) 木村文則，前田 亮，吉川正俊，植村俊亮：ディレクトリ型検索エンジンを利用した言語横断情報検索，第 1 回情報科学技術フォーラム論文集，第 2 分冊，pp.69–70 (2002).
3) Kimura, F., Maeda, A., Yoshikawa, M. and Uemura, S.: Cross-Language Information Retrieval using Web Directories, *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing* (*PACRIM '03*), pp.911–914 (2003).
4) 酒井哲也，梶浦正浩，住田一男，Jones, G., Collier, N.：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，情報処理学会論文誌，Vol.40, No.11, pp.4075–4086 (1999).
5) Jansen, B.J., Spink A. and Saracevic, T.: Real life, real users and real needs: a study and analysis of user queries on the Web, *Information Processing & Management*, Vol.36, No.2, pp.207–227 (2000).
6) Hull, D.A.: Using structured queries for disambiguation in cross-language information retrieval, *Electronic Working Notes of the AAAI Symposium on Cross-Language Text and Speech Retrieval* (1997).
7) 奥村明俊，石川 開，佐藤研治：コンパラブルコーパスと対訳辞書による日英クロス言語検索，自然言語処理，Vol.5, No.4, pp.77–93 (1998).
8) Lin, C.J., Lin, W.C., Bian, G.W. and Chen, H.H.: Description of the NTU Japanese-English cross-lingual information retrieval system used for NTCIR workshop, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.145–148

(1999).
9) 木村文則，前田 亮，吉川正俊，植村俊亮：Webディレクトリを利用した言語横断情報検索における特徴語抽出，「情報アクセスのためのテキスト処理」シンポジウム 発表論文集，pp.1–8 (2003).
10) Kimura, F., Maeda, A., Yoshikawa, M. and Uemura, S.: Cross-Language Information Retrieval Based on Category Matching Between Language Versions of a Web Directory, *Proc. 6th International Workshop on Information Retrieval with Asian Languages* (*IRAL2003*), pp.153–159 (2003).
11) Frakes, W. and Baeza-Yates, R.: *Information Retrieval: Data Structures and Algorithms*, chapter 7, Prentice-Hall (1992).
12) Chen, K-H., Chen, H-H., Kando, N., Kuriyama, K., Lee, S., Myaeng, S.H., Kishida, K., Eguchi, K. and Kim, H.: Overview of CLIR Task at the Third NTCIR Workshop, *Working Notes of the 3rd NTCIR Workshop Meeting*, pp.1–38 (2002).

(平成 15 年 12 月 20 日受付)
(平成 16 年 4 月 6 日採録)

(担当編集委員 関根 純)

木村 文則（学生会員）

2001 年大阪教育大学大学院教育学研究科総合基礎科学専攻修士課程修了．同年より奈良先端科学技術大学院大学情報科学研究科情報システム学専攻博士後期課程に在学中．言語横断情報検索の研究に従事．

前田 亮（正会員）

1995 年図書館情報大学図書館情報学部卒業．1997 年同大学大学院図書館情報学研究科修士課程修了．2000 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．日本学術振興会特別研究員，科学技術振興事業団 CREST 研究員を経て，2002 年立命館大学理工学部助教授，2004 年より同大学情報理工学部助教授，現在に至る．2000 年〜2001 年米国バージニア工科大学客員研究員．ディジタル図書館，多言語情報処理，情報検索等に興味を持つ．平成 10 年度情報処理学会論文賞受賞．ACM，電子情報通信学会，情報メディア学会各会員．

宮崎 純（正会員）

1992 年東京工業大学工学部情報工学科卒業．1997 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（情報科学）．同大学助手を経て，2003 年より奈良先端科学技術大学院大学情報科学研究科助教授．2000 年〜2001 年テキサス大学アーリントン校客員研究員．高性能・高機能データベースシステムの研究に従事．電子情報通信学会，IEEE CS，ACM SIGMOD 各会員．

吉川 正俊（正会員）

1980 年京都大学工学部情報工学科卒業．1985 年同大学大学院工学研究科博士後期課程修了．工学博士．京都産業大学，奈良先端科学技術大学院大学を経て，2002 年から名古屋大学情報連携基盤センター教授．The VLDB Journal および Information Systems（Elsevier/Pergamon）の編集委員．XML データベース，多次元空間索引等の研究に従事．

植村 俊亮（フェロー）

1964 年京都大学工学部電子工学科卒業．1966 年同大学大学院工学研究科修士課程修了．同年電気試験所（産業技術総合研究所）．1970 年マサチューセッツ工科大学電子システム研究所客員研究員，1981 年電総研ソフトウェア部プログラム研究室長，1988 年東京農工大学教授を経て，1993 年から奈良先端科学技術大学院大学情報科学研究科教授．データ工学，データベースシステムの研究に従事．工学博士．IEEE Fellow，電子情報通信学会フェロー．現在，情報処理学会理事，日本情報考古学会理事，データベース振興センター評議員等．