D-045

言語横断情報検索におけるWebディレクトリを利用した訳語の曖昧性解消
Translation Disambiguation Using Web Directory for CLIR

木村 文則† 前田 亮‡ 宮崎 純† 吉川 正俊§ 植村 俊亮†
Fuminori Kimura Akira Maeda Jun Miyazaki Masatoshi Yoshikawa Shunsuke Uemura

1. まえがき

WWWが世界的に普及し，母国語以外の文書も電子的に入手することが容易になった．しかし，母国語以外の言語で検索することは利用者にとって負担となる．それゆえ，このような文書を利用者の母国語で検索できることが望まれる．そこで，ある言語で書かれた文書群を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）に関する研究が盛んになっている．例えば，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ている．しかしこの手法では，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．

そこで我々は，言語横断情報検索において，例えばYahooのようなWebディレクトリを言語資源として利用する手法を提案している．この手法では，問合せが適合するカテゴリを推定し，そのドメインに対して適切な訳語を推定することで，翻訳された問合せの曖昧性解消を行う．本論文では，問合せが適合するカテゴリの選択方法について提案手法を改良し，その評価を行った．

2. 関連研究

CLIRにおいて，問合せを翻訳する場合に対訳辞書を用いることが多いが，このとき訳語の曖昧性解消が問題となる．その解決方法として，コーパスを用いる手法が研究されている．しかし，このような手法は，検索要求とコーパス間のドメインの相違による検索性能への影響が指摘されている [1].

Web検索などでは多様なドメインの検索要求への対応が求められるが，それぞれのドメインについて対応するコーパスを用意するのは現実的ではない．そこで我々は，Yahooなどの Webディレクトリを言語資源としてCLIRに適用する手法を提案している [2]．Webを言語資源として利用した研究として，コンパラブルコーパスの自動収集にWebを利用する手法 [3] などがある．しかし，我々の手法は，問合せの翻訳においてWebディレクトリを利用する点で，これらの研究とは異なっている．

3. 提案手法

図1は提案手法のシステムの概要および検索の流れを表している．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，各言語版の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．

†奈良先端科学技術大学院大学 情報科学研究科
‡立命館大学 情報理工学部 メディア情報学科
§名古屋大学 情報連携基盤センター

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．

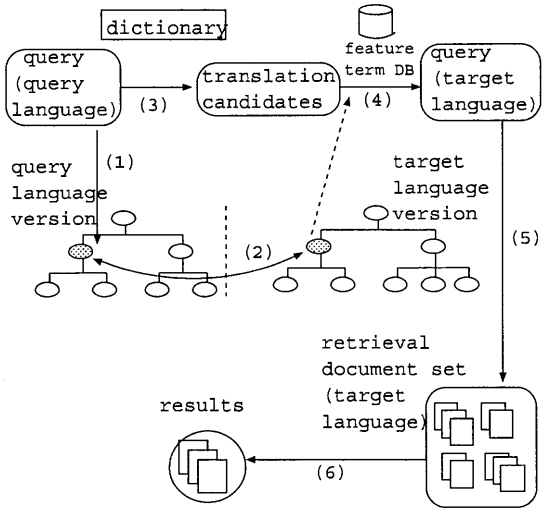

図 1: 提案システムの概要および検索の流れ

3.1 前処理

前処理と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ディレクトリの全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
   全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．

言語間でのカテゴリの対応付けは，どのような方法で行ってもよい．例として，手動で行う，カテゴリの特徴語を比較するなどの方法が考えられる．

3.2 検索処理

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全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．そのうちで，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．
(2) 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．
(3) 対訳辞書を用いて，問合せ語の訳語候補をすべて抽出する．(詳細は 3.2.1)
(4) 選択された対応カテゴリの特徴語集合を利用して，最適な訳語を決定する．(詳細は 3.2.1)
(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．
(6) 検索結果を得る．

3.2.1 問合せの翻訳

本節では，3.2 で述べた，検索処理における問合せの翻訳の手法（3.2 の検索における処理手順の (3)(4)）について詳しく述べる．まず，問合せ中の各単語に対する対訳辞書の全ての訳語を，訳語の候補として抽出する．抽出された全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”）の特徴語に含まれているかを調べる．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，例えば，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良い場合もある．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．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みなす．

以前の研究 [2] では，一つの対応カテゴリの特徴語集合のみについてしか調べなかったため，訳語が得られないことも多かった．そこで本論文では，そのような場合，2 番目以降の適合カテゴリについても調べるように問合せ翻訳の手法を改良した．対応カテゴリの特徴語集合に含まれているうちの重みが最も大きい訳語候補を選択するというのは変わらないが，最初に調べたカテゴリに訳語候補が一つも存在していない場合は，その次に適合度の高い対応カテゴリの特徴語集合も同様に調べるように変更し，最大三つのカテゴリについて調べるように改良した．これにより，訳語が得られない可能性を減少させることができる．

4. 実験

提案手法の有効性を検証するため，NTCIR-3 CLIR タスクのテストコレクションを用いて検索の実験を行った．問合せには日本語問合せの TITLE フィールドを用い，これを提案手法により英語に翻訳し，英語の文書群に対して検索を行った．

50 件の問合せのうち，本論文の改良版手法により新たに訳語が追加されたのは 6 件であった．これらの問合せのうち，1 件は平均適合率に変化がなかったが，3 件で向上した．なお，残り 2 件はもともと適合文書が存在しないため，本テストコレクションでは評価されていない．

図 2 は，実験結果を適合率・再現率グラフに表したものである．“TI-org” が改良前の手法 [2]，“TI-rev” が本論文で提案した手法である．わずかではあるが，改良版のほうが以前の手法よりも上回っている．11 点平均適合率においても，“TI-org” が 0.0803，“TI-rev” が 0.0829 と，0.26 ポイント上回った．

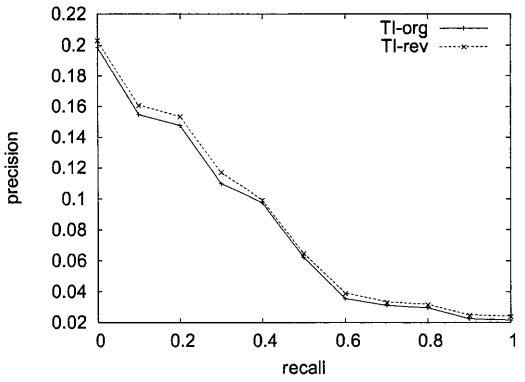

図 2: 検索結果の適合率・再現率グラフ

5. おわりに

我々は，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ているが，本論文では，検索精度を向上させるため，問合せ翻訳手法の改良を行った．また，本手法の有効性を検証するために検索の実験を行い，検索精度が向上することを示した．

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挙げられる．

参考文献

[1] 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, October 1998.

[2] 木村文則, 前田亮, 宮崎純, 吉川正俊, 植村俊亮. 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. 情報処理学会論文誌:データベース, Vol. 45, No. SIG 7(TOD 22), pp.208-217, June 2004.

[3] W. Kraaij, J.Y. Nie, and M. Simard. Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval. *Computational Linguistics*, Vol. 29, No. 3, pp.381-419, September 2003.