多言語検索における言語横断キーワード抽出システムの構築

浦江 宏志† 手塚 太郎† 木村 文則† 前田 亮†
立命館大学情報理工学部†

概要

検索クエリと検索結果の単純な翻訳に基づく既存の多言語検索システムを使用した場合，検索結果の要約(スニペット)に意味のある情報が少ししか含まれないという問題がある．また，検索結果のWebページを開き，それが検索要求に合致しているかを確認するために必要な時間も，ユーザの母国語で書かれたWebページに比べて長くなってしまう．そこで，多言語検索で検索結果として得られるWebページから，各Webページのコンテンツを特徴づけるキーワードを抽出し，翻訳を行う．これらのキーワードを検索結果と合わせて提示することで，検索の段階で各Webページに必要な情報が載っているかをより確実に判断するシステムを開発した．

1. はじめに

検索方法の一つに言語横断検索[1]がある．ユーザが母国語で入力したクエリを指定された他言語に翻訳し，新たなクエリとする．その新たなクエリを用いて，他言語で書かれたサイトを検索する検索方法である．これにより，ユーザは他言語に対する知識がなくとも，他言語で書かれたWebページを検索することができる．

現在，Google や Yahoo!などの検索サービスでは，言語横断検索を用いた多言語検索が利用できる．これにより，ユーザは母国語のみで多言語で書かれた Web ページを検索できるようになったが，使い勝手は良いとは言えない．多言語検索は検索結果がユーザの母国語とは異なる言語になるため，検索結果を翻訳して表示する必要がある．しかし，既存の多言語検索サービスでは検索結果に表示されるタイトルやスニペットの部分のみを翻訳しているため，翻訳された検索結果から得られる情報は少ない．そのため，検索結果からでは各 Web ページのコンテンツを把握しにくく，目的の情報が載っているかを判断するためには，各 Web ページにアクセスしなければならない．この作業は，母国語で書かれた Web ページを見るよりもはるかに時間がかかる．

本論文では，多言語検索において目的の情報が，素早く，確実に得られないという問題を解決するために，多言語検索における言語横断キーワード抽出システムを用いた解決法を提案し，その有効性を検証した．

Cross-Language Keywords Abstraction System for Multi-Lingual Information Retrieval
†Hiroshi Urae, Taro Teduka, Huminori Kimura, Akira Maeda
College of Information Science and Engineering, Ritsumeikan University

2. 提案手法

本論文で提案する手法では，検索結果に各Webページのコンテンツを具体的に表すようなキーワードを付け加えることで，検索結果の情報量を増やす．これにより，検索結果の段階で各Webページのコンテンツがイメージしやすくなり，余計なWebページへのアクセスを減らすことができる．今回実装したシステムは英語を母国語とするユーザ向けのものであり，検索対象は日本語で書かれているWebページである．システムの概要図を以下に示す．

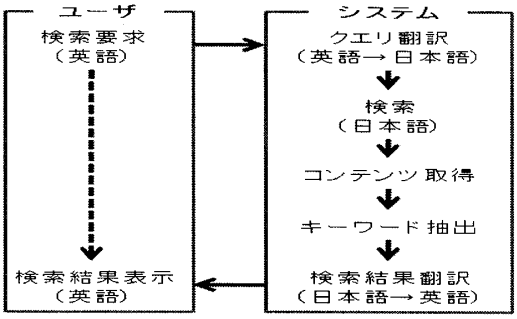

図1　システムの概要図

図1で示したように，システムは次の5工程に分けられる．

2.1. クエリ翻訳

ユーザが英語で入力したクエリを日本語に機械翻訳する．

2.2. 検索

日本語に翻訳したクエリで，日本語で書かれたWebページを検索する．

2.3. コンテンツ取得

検索結果として得られた Web ページにアクセスし，HTML ソースを取得する．取得した HTML ソースから HTML タグなどの不要な情報を削除し，Web ページのコンテンツ部分のみを

取得する．

2.4. キーワード抽出

抽出したコンテンツを日本語の形態素解析ツールであるMeCab‡にかけキーワード候補を抽出する．キーワード候補は「数」，「接尾」，「代名詞」，「非自立」，「副詞可能」以外と解析された名詞のうち，読みがあるものとしている．また，MeCabのデフォルトの辞書だけでは解析できないような複合語が多く存在するので，Wikipediaのタイトルを辞書に追加し，これらもキーワード候補としている．キーワード候補はWebページ毎に出現回数を数え，各キーワード候補の出現回数をtf(Term Frequency)とする．さらに事前に用意したidf(Inverse Document Frequency)を用い各キーワード候補に対しtf-idfを算出する．idfは大量の文書から作成するため，文書のコンテンツが偏る場合がある．この問題を防ぐため，情報が分野ごとにわけられているYahoo! Japanの各カテゴリから，1サイトずつ登録サイトのコンテンツ(全部で約18万サイト)を取得し，同じようにMeCabを用いて形態素解析をして作成した．Webページ毎に，tf-idfの高いキーワード候補上位5件を選び，そのWebページのキーワードとする．

2.5. 検索結果翻訳

各Webページのタイトル，スニペット，キーワードを機械翻訳した新たな検索結果(図2)を作成し，ユーザに提示する．

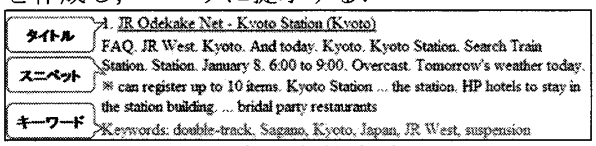

図2　新たな検索結果

3. 評価実験

今回の評価実験では，従来の多言語検索での検索結果に対し，キーワードを付加したことによって，より各Webページのコンテンツが把握しやすくなったかを人手により評価する．評価実験に用いたWebページは50件である．まず従来の検索結果をWebページ毎に3段階評価する．十分にコンテンツを把握できる場合は1，把握しにくい場合は0，全くわからない場合は-1のスコアを付ける．次に，キーワードを付加した提案手法の検索結果に対して，同様に3段階で評価する．その後，キーワードを付加した提案手法の検索結果のスコアから従来の検索結果のスコアを引き，変化量(-2～2の5段階)を算出する．結果は表1である．

今回の実験では，従来の検索結果より，提案手法の検索結果のほうがコンテンツをより把握できた(変化量が正)となった結果は23件であった．一方，提案手法の方が把握しづらくなった(変化量が負)結果は2件であった．

表1　実験結果

従来の検索結果のスコア	提案手法の検索結果のスコア	変化量	件数
0	-1	-1	1
1	0	-1	1
-1	-1	0	3
0	0	0	11
1	1	0	11
-1	**0**	**1**	**3**
0	**1**	**1**	**14**
-1	**1**	**2**	**6**

4. 考察

実験結果から，半数近くの検索結果において，提案手法は多言語検索の検索結果からWebページのコンテンツを把握しやすくすると認められた．特に，スニペットが途中で途切れていたり，句読点なしで書かれている場合は，スニペットが正しく翻訳されにくいため，提案手法が有効であった．

一方で，Wikipediaなど Webページのコンテンツが Webサイトの特性によって把握出来てしまう場合や，コンテンツが動的に変化する場合は提案手法があまり有効ではなかった．

5. まとめ

本論文では，多言語検索において目的の情報が，素早く，確実に得られないという問題を解決するために，多言語検索における言語横断キーワード抽出システムを用いた解決法を提案し，評価実験によりその有効性を示した．本研究の今後の課題としては，キーワードの抽出・翻訳精度の向上，多言語化の拡大が挙げられる．

参考文献

[1]木村 文則，前田 亮，波多野 賢治，宮崎純，植村 俊亮：Webディレクトリの階層構造を利用した検索対象文書の分野推定に基づい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.49，pp.59-71（2008）

‡ http://mecab.sourceforge.net/