DEIM Forum 2010 D10-1

可視化によるバイリンガル検索支援

内藤宗一朗† 太田 学††

† 岡山大学工学部情報工学科 〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 岡山大学大学院自然科学研究科 〒 700–8530 岡山県岡山市北区津島中 3–1–1
E-mail: †,††{naito,ohta}@de.cs.okayama-u.ac.jp

あらまし Web サイトの検索を行う際に，日本語だけでなく英語のサイトも検索対象に含められれば，より多くの情報が得られる．しかし，母国語でない言語で検索を行い，さらにその結果の内容を把握することは困難である．本稿では，入力されたクエリを基に日英二言語で検索を行い，結果を可視化することで内容把握を容易にする手法を提案する．提案システムでは，ユーザが入力した日本語クエリから自動的に英語クエリを生成し，日英二言語で検索を行う．またその結果から特徴語を抽出し，それらの関係を可視化することで検索結果の内容把握を支援する．実験では，生成する英語クエリを評価し，可視化した検索結果について分析した．
キーワード 検索支援，データ可視化，言語横断検索

Bilingual Retrieval Support Using Visualization

Soichiro NAITO† and Manabu OHTA††

† Department of Information Technology, Faculty of Engineering, Okayama University
3–1–1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama, 700–8530 Japan
†† Graduate School of Natural Science and Technology, Okayama University
3–1–1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama, 700–8530 Japan
E-mail: †,††{naito,ohta}@de.cs.okayama-u.ac.jp

Abstract We can get more information if we simultaneously search English Web sites using an English query when searching Japanese sites using a Japanese query. However, it is sometimes difficult for ordinary Japanese people to hit upon good English keywords and to understand English search results. This paper proposes a bilingual retrieval support system which enables us to search both Japanese and English sites simultaneously and to understand the visualized search results easily. The proposed system automatically generates an English query by translating a Japanese one inputted by a user and simultaneously searches Japanese and English sites. In addition, the system extracts feature terms from the search results and visualizes them using the feature terms. We evaluated the translated English queries and analyzed the visualized search results.
Key words Retrieval Support, Data Visualization, Cross-language Retrieval

1. はじめに

インターネット上の情報を入手する手段の一つとして，サーチエンジンを利用する機会が増えている．その際，より広い範囲から情報を収集するためには日本語のサイトを検索するだけでは不十分なことも多い．その場合には，他の言語のサイトも検索対象に含める必要がある．しかし，母国語でない言語でクエリを入力し，その検索結果の内容を把握することは困難である．

そこで本研究では，ユーザが入力したクエリを基に日英二言語で検索を行い，その結果を可視化してユーザに提示する．入力された日本語クエリから英語クエリを自動生成し，英語における検索を支援する．また，得られた検索結果から特徴語を抽出し，それを用いて結果を可視化したものをユーザに提示することで，容易に検索結果の内容を把握できるようにする．

本稿では，2 節で関連研究について述べる．3 節で提案システムについて説明し，4 節でシステムの実装および動作例について述べる．5 節で評価実験と考察を行い，6 節でまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 言語横断検索

樋口ら [1] は特許情報を対象とした言語横断検索システムを開発した．このシステムでは対訳辞書を用いてクエリを翻訳し，特許データベースを検索する．本稿の提案システムではクエリの翻訳に対訳辞書を用いず，Web 資源である Wikipedia [2] と Google AJAX Language API [3] を用いる．

中崎ら [4] は，ある同一のトピックについての記述があるブログサイトを日英各言語で検索し，その内容を二言語間で対照分析するシステムを実装した．このシステムでは Wikipedia の日英二言語エントリからトピックの日英二言語表現を取得し，検索を行っている．本研究でもクエリの翻訳に Wikipedia を利用するが，本研究では検索対象をブログサイトに限定しない．

新井ら [5] は多言語情報へのアクセス支援を目的として，Wikipedia の言語間リンクを利用した多言語対訳システムを構築した．彼らは Wikipedia の言語間リンクの訳語としての有用性を評価し，Web 検索においても多言語アクセス支援が可能であることを示した．このシステムと本稿の提案システムでは，Wikipedia の言語間リンクを用いて訳語を取得し，それを用いて複数言語で検索を行う点が一致している．しかし本稿の提案システムでは，複数言語で検索を行い結果を取得するだけでなく，その検索結果を可視化して内容把握の支援を行う．

2.2 可視化による検索支援

結城 [6] はベン図を用いた可視化による検索支援システムを提案した．このシステムではベン図上で円をクエリに対応させ，その円の大きさで各クエリによる検索ヒット数を表現する．また円同士の重なりにより，各クエリによる AND,OR,NOT 検索のヒット数を可視化した．本稿の提案システムでは，検索のヒット件数ではなく結果の概要を可視化する．

長畑ら [7] は，検索結果中の特徴語とそれらの関係を抽出し，可視化することで検索支援を行うシステムを作成した．このシステムでは，連続した検索により得られる検索結果から特徴語の出現頻度の変化を求め，各特徴語の出現頻度の変化と特徴語間における変化の類似度を基に可視化を行う．また，各特徴語間の共起度に基づく可視化も行っている．可視化には特徴語をノード，特徴語間の関係をエッジとしたバネモデルを用いている．本稿では長畑らの可視化手法を利用して，日本語クエリのみを入力とする日英二言語検索と，その両言語の検索結果を統合して可視化する手法を提案する．

3. 提案システム

3.1 システムの概要

本研究は，日英同時検索と，検索結果の可視化による内容把握支援の二つを目的とする．よって提案手法ではまず，ユーザが入力した日本語クエリから自動的に英語クエリを生成し，それら二言語のクエリで日英同時検索を行う．そして得られた検索結果から特徴語やその出現頻度といった情報を抽出し，それを可視化して提示することで結果の内容把握を支援する．

3.2 バイリンガル検索

入力された日本語クエリを基に日英二言語で検索を行う方法について述べる．まず入力された日本語クエリを翻訳し，英語クエリを生成する．クエリの英訳には Wikipedia の言語間リンクと Google AJAX Language API [3] を用いる．そして日本語クエリと生成した英語クエリを用いて検索を行う．なお検索には Yahoo!ウェブ検索 Web API [8] を用いる．

3.2.1 クエリの英訳

まず，日本語クエリを翻訳し英語クエリを生成する方法について述べる．クエリはそのまま翻訳するのではなく，スペースによって区切られた検索語をそれぞれ翻訳していく．

英訳のため，本システムはまず Wikipedia の言語間リンクを用いる．同一の事柄について書かれた記事が複数の言語で存在する場合，Wikipedia はそれらの記事間のリンクを言語間リンクとして記事中に保持している．本システムでは，この言語間リンクによって結ばれた日英の記事のタイトルを対訳辞書として利用する．英訳を行う検索語で日本語 Wikipedia を検索し，該当する記事の言語間リンクから英語記事のタイトルを取得し，それを英訳とする．検索語によっては「曖昧さ回避のためのページ」がヒットしたり，他の記事にリダイレクトされる場合がある.「曖昧さ回避のためのページ」はページ中に言語間リンクが存在するため，他の記事と同様に語の英訳が可能である．他の記事にリダイレクトされる場合は，リダイレクト先の記事で言語間リンクを参照し英訳を行う．

また本システムは，英訳に Google AJAX Language API を利用する．Google AJAX Language API は Google [9] が提供する Web API の一つであり，テキストを指定した言語に翻訳する機能を持っている．日本語を入力とし，翻訳先の言語に英語を指定することでクエリの英訳を取得することができる．

Wikipedia は百科事典であるため，多くの専門用語や固有名詞についての記事が存在する．そのような専門用語や固有名詞については Wikipedia による翻訳の方が良い英訳を得られる可能性が高い．その反面，Wikipedia 中に記事が存在しない場合や言語間リンクが取得できない場合は英訳を行うことができない．そのため，提案システムの英訳はまず Wikipedia を用いた方法を試み，それで英訳を取得できなかった場合に Google AJAX Language API により英訳する．

3.2.2 二言語のクエリによる検索

入力された日本語クエリと生成した英語クエリの二つを用いて検索を行う．Yahoo!ウェブ検索 Web API は検索する言語の指定ができるため，それぞれのクエリにそれぞれの言語を指定して検索を行う．取得した検索結果から後述する特徴語とその頻度情報等を抽出する．

3.3 特徴語抽出

取得した検索結果から特徴語を抽出する方法について述べる．特徴語は日英それぞれの検索結果から抽出し，出現頻度に基づくスコア付けを行う．

3.3.1 日本語検索結果からの特徴語抽出

日本語の検索結果からの特徴語抽出は，長畑ら [7] の方法を参考にした．

まず形態素解析器 Sen [10] を用いて検索結果から形態素を抽出する．抽出された形態素のうち，以下のものを特徴語の構成要素とする．

(1) 名詞あるいはカタカナのみで構成される形態素
(2) 英数字のみで構成される形態素
(3) 接頭辞
(4) 区切り記号「・」,「—」,「.」
(5) 連体助詞「の」

これらの形態素を連結し，特徴語を生成する．形態素の連結は，日本語の特徴語と日本語以外の特徴語に分けて行う．(1), (3)，(4)，(5) が連続する場合は日本語の特徴語として連結する．(2)，(4) が連続する場合は外国語の特徴語として連結する．ただし，(4) は特徴語の先頭には現れないものとする．

また，形態素の連結により生成された特徴語は長くなりすぎる場合がある．これを防ぐため，接頭辞の前および接尾辞の後ろで特徴語を分割する．

3.3.2 英語検索結果からの特徴語抽出

英語の検索結果からの特徴語抽出は，TermExtract [11] で用いられている方法を参考にした．まず Monty Tagger [12] を用いて検索結果を単語に分割し，品詞タグを付与する．このとき特徴語は以下の単語を構成要素とする．

(1) 名詞，固有名詞
(2) 外国語
(3) 基数
(4) 形容詞
(5) 所有格語尾
(6) of
(7) 過去分詞形の動詞

そしてこれらの単語を連結し，特徴語を生成する．各単語は以下に示す規則に従い連結する．

- 名詞 $\cdots$ 名詞，固有名詞，形容詞，基数，過去分詞形の動詞の後ろに連結する．
- 外国語 $\cdots$ 単体で特徴語とする．
- 基数 $\cdots$ 他の単語の後ろには連結しない．
- 形容詞 $\cdots$ 形容詞，所有格語尾，基数の後ろに連結する．
- 所有格語尾 $\cdots$ 名詞，固有名詞の後ろに連結する．
- of $\cdots$ 名詞，固有名詞の後ろに連結する．
- 過去分詞形の動詞 $\cdots$ 他の単語の後ろには連結しない．

所有格語尾と of 以外の単語は特徴語の先頭となることができる．特徴語は名詞，固有名詞，外国語のいずれかで終わるものとする．固有名詞以外の名詞については語の先頭を小文字にし，複数形の場合は単数形にする．また，特徴語が長くなりすぎることを防ぐため，特徴語を構成する単語の数は 3 以下とする．

3.3.3 特徴語のスコアと類似度

抽出した特徴語について，そのスコアと類似度の算出を行う．特徴語のスコアおよび類似度の定義は長畑ら [7] の方法を参考にした．はじめに，式 (1) で特徴語のスコアを定義する．ここで $tf\_t_i^k$ は k 回目の検索において特徴語 w_i が検索結果のタイトル中に出現した回数，$tf\_s_i^k$ は k 回目の検索において特徴語 w_i が検索結果のサマリ中に出現した回数である．$weight\_t$ と $weight\_s$ は重みであり $weight\_t = 8$，$weight\_s = 7$ とする．この式は，出現頻度が高い特徴語には大きいスコアを割り当て，またタイトル中に出現する特徴語はサマリ中に出現する特徴語より重みが大きい．

$$score_i^k = \begin{cases} 0 & \text{if } k = 0 \\ weight\_t * tf\_t_i^k & \\ \quad + weight\_s * tf\_s_i^k & \text{otherwise} \end{cases} \tag{1}$$

次に，特徴語間の類似度について述べる．本稿では，特徴語の出現頻度とその推移を基に特徴語間の類似度を算出する．出現頻度の差が小さい特徴語，連続する検索において出現頻度の変化が似ている特徴語は類似度が大きくなる．すなわち特徴語 w_i と w_j の類似度 $sim_{i,j}$ を以下の式で定義する．

$$sim_{i,j} = \sum_{k=1}^{count} \frac{k}{count} \cdot \left(diff_{max}^k - diff_{i,j}^k \right) \tag{2}$$

$$diff_{max}^k = \max_{i,j} diff_{i,j}^k \tag{3}$$

$$\begin{aligned} diff_{i,j}^k =& |sub_i^k - sub_j^k| + |score_i^k - score_j^k| \\ & + |score_i^{k-1} - score_j^{k-1}| \end{aligned} \tag{4}$$

$$sub_i^k = score_i^k - score_i^{k-1} \tag{5}$$

式 (2) の $count$ は検索回数である．また式 (2) では，直近の検索結果を重視するために分子を k 倍している．式 (4) は k 回目の検索における特徴語 w_i と w_j のスコア等の差異の大きさであり，これが小さいほど w_i と w_j は類似しているとみなす．すなわち $diff_{i,j}^k$ は，今回と前回の検索結果におけるスコアの差と，前回から今回までのスコアの推移の差を基に算出する．式 (5) は，前回の検索結果と今回の検索結果において比較した特徴語 w_i のスコアの推移である．

3.4 検索結果の可視化

特徴語を利用した検索結果の可視化について述べる．特徴語をノード，特徴語間の関係をエッジとしたグラフをユーザに提示し，検索結果の内容把握を支援する．出現頻度の推移に基づく可視化と共起度による可視化をそれぞれ言語別に行い，さらに日英両方の結果を統合した可視化を行う．また，バネモデル [7] によるノードの再配置を行い，グラフの可読性を向上させる．

3.4.1 出現頻度の推移に基づく可視化

抽出した特徴語について，出現頻度の推移および推移の類似度に基づき可視化を行う．連続した検索における結果とその推移を可視化することで，検索結果の概要およびその変化の把握を支援する．可視化では抽出した特徴語を全て提示すると結果の可読性が低下するので，提示する特徴語の数を制限する．そのため以下に示す $change_i$ と $theme_i$ を定義する．

$$change_i = \frac{1}{2}|sub_i^k| \tag{6}$$

$$theme_i = \frac{1}{count}\sum_{k=1}^{count} score_i^k \tag{7}$$

式 (6) は前回の検索と今回の検索における特徴語 w_i のスコアの差の絶対値の 1/2 であり，出現頻度の変化が大きいほど $change_i$ の値も大きくなる．式 (7) は今回の検索までの特徴語 w_i のスコアの平均であり，連続した検索で高い出現頻度を維持しているほど $theme_i$ の値も大きくなる．このように定義した $change_i$ と $theme_i$ のいずれか大きい値を，可視化における特徴語 w_i の優先度とする．その優先度が高い特徴語から順に可視化してユーザに提示する．また，出現頻度の推移をユーザに提示するため，その推移のパターンによってノードの色分けを行う．$change_i \leqq theme_i$ となる特徴語は，出現頻度を高い水準で維持しているとして黄色のノードで表示する．$change_i > theme_i$ となる特徴語については，$0 \leqq sub_i^k$ となるものは出現頻度が増加した特徴語として白色のノードで，$0 > sub_i^k$ となるものは出現頻度が減少した特徴語として灰色のノードで表示する．また出現頻度の推移の似ている特徴語間にはエッジを生成し，その類似度をエッジの色の濃淡で表現する．類似度が高い特徴語間はより濃い色のエッジで結ばれる．

3.4.2 共起度に基づく可視化

特徴語のスコアと共起度に基づいて可視化を行う．検索結果において同じタイトル，サマリ内に共起する語は強い関連を持つと考えられる．そこで特徴語の共起関係に基づき可視化することで，特徴語間の関連の強さをユーザに提示する．ここで，同一のタイトルおよびサマリにおいて特徴語 w_i と w_j が出現しているとき，特徴語 w_i と w_j は共起しているとする．特徴語 w_i と w_j の関連の強さを表す共起度 $coo_{i,j}$ を以下の式により定義する．

$$coo_{i,j} = \frac{1}{N}\sum_{d=1}^{N} co\_occur_{i,j}^d \tag{8}$$

$$co\_occur_{i,j}^d = \begin{cases} 1 & \text{if } w_i \in D_d \ \wedge w_j \ \in D_d \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

ここで N は取得した検索結果の文書の総数である．$co\_occur_{i,j}^d$ は d 番目の文書のタイトルおよびサマリ D_d において，特徴語 w_i と w_j が共起する場合に 1 となる．出現頻度の推移に基づく可視化の場合と同様に，特徴語を全て提示するのは可読性の点で有効でないため，ここでは式 (1) のスコアが高い語を優先的に提示する．特徴語間における共起度の強さは，前節と同様にエッジの濃淡で表現する．また，クエリに含まれる検索語は他の多くの特徴語と共起関係にあると予想できるため，グラフ中に表示しない．

3.4.3 特徴語の統合

同じような内容の特徴語を複数提示しても内容把握には有効でないため，そのような特徴語は一語に統合する．特徴語 w_x を部分文字列として含む特徴語 w_y が存在する場合，w_x を w_y に統合する．統合後の w_y が持つスコア，他の特徴語との類似度は w_x と w_y が持っていたそれらの平均とする．日本語の場合，統合される特徴語 w_x は形態素一つからなる特徴語に限定し，統合先となる特徴語 w_y は形態素二つからなる特徴語に限定する．英語の場合，統合される特徴語 w_x は単語一つからなる特徴語に限定し，統合先となる特徴語 w_y は単語二つからなる特徴語に限定する．

3.4.4 バイリンガル検索結果の可視化

日本語と英語の両方の検索結果を考慮して，特徴語の出現頻度とその推移に基づく可視化を行う．日英両方の結果を統合して可視化することで，二言語で行った検索の結果の概要を同時に閲覧することができる．また，対訳関係にある特徴語ノードを統合して可視化するため，日英の検索結果の共通点や相違点を読み取ることができる．可視化は 3.4.1 項と同様の方法で行う．日英両言語の特徴語をあわせて可視化するため，提示する特徴語は単言語の場合よりも通常多くなる．

日英二言語の特徴語の可視化では，対訳関係にある特徴語ノードを一つに集約するため，提示する日本語の特徴語を英訳し，それと一致する特徴語が英語の方に存在すれば，これらのノードを統合する．統合後のノードは“ 日本語（英語）”という形式にし，日英両言語の検索結果の特徴語として抽出されたことが分かるよう表示する．また，統合後のノードのスコアとその変化，他のノードとの類似度は元の二つのノードの平均とする．

4. 実　　装

提案システムのプロトタイプを実装した．システムのインタフェース画面を図 1 に示す．本例は“ イチロー ”で検索した後,“ イチロー　シアトルマリナーズ ”で検索したときのものである．

画面の左上にはクエリ入力欄や検索ボタンなどを配置した．画面の左下には，得られた検索結果を可視化したグラフが表示される．画面の右側にはタイトル，URL，サマリから構成される検索結果が表示される．

5. 評価実験と考察

5.1 クエリ英訳の評価

入力された日本語クエリを提案手法が適切な英語クエリに翻訳できるか実験により評価した．

5.1.1 評 価 実 験

四つのパターンの日本語クエリについて，適切な英語クエリを生成できるか実験を行った．一つ目は名詞，固有名詞からなるクエリである．二つ目は多義語であり，これらは Wikipedia では「曖昧さ回避」[13] の対象となる．Wikipedia における「曖昧さ回避」とは，内容が異なるのに名前が同じ記事について，それらを判別しやすくする仕組みのことである．三つ目は略称

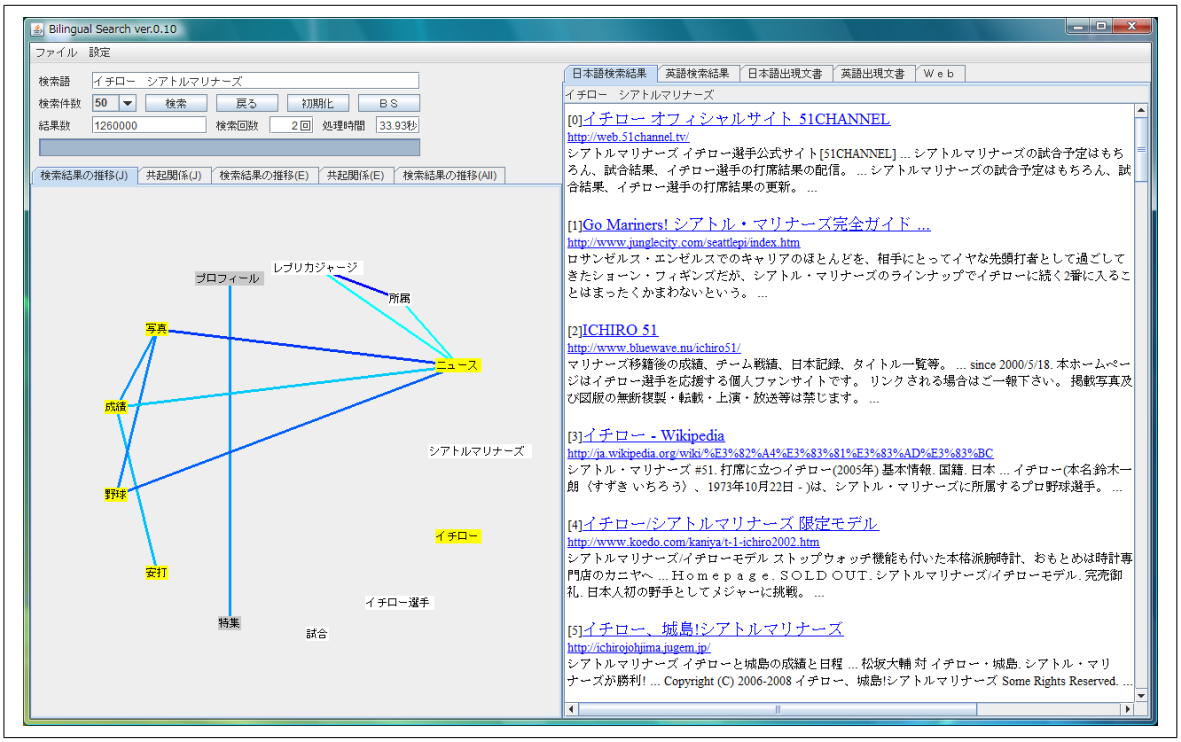


図 1　インタフェース画面

表 1　クエリの英訳の評価 (1)

日本語クエリ	生成される英語クエリ
普天間飛行場 / 移設	Marine Corps Air Station Futenma / Relocation
バラク・オバマ / 支持率	Barack Obama / Approval
ハリー・ポッター / 映画	Harry Potter / Film

表 2　クエリの英訳の評価 (2)

日本語クエリ	生成される英語クエリ
アバター	Avatar
オリックス	Orix

表 3　クエリの英訳の評価 (3)

日本語クエリ	生成される英語クエリ
普天間基地	Marine Corps Air Station Futenma
オバマ大統領	Barack Obama
キング牧師	Martin Luther King, Jr.
パソコン	Personal computer

表 4　クエリの英訳の評価 (4)

日本語クエリ	生成される英語クエリ
アボカド / 調理 / 方法	Avocado / Cooking / Way
コーヒー / 入れ方	Coffee / You put

や別称など，Wikipedia においてリダイレクトの対象となる語である．四つ目は物事の方法を調べるときによく使われるクエリである．

表 1 は名詞，固有名詞からなる日本語クエリを翻訳して英語クエリを生成した結果である．“ 普天間飛行場 ”や“ バラク・オバマ ”,“ ハリー・ポッター ”のような通常の和英辞典には載っていない固有名詞に対しても，Wikipedia を利用することで適切に英語クエリを生成できていた．また，”移設”や“ 支持率 ”,“ 映画 ”のような Wikipedia に記事がない普通名詞については，Google AJAX Language API によりおおむね適切に翻訳されている．

表 2 は多義語を日本語クエリとして，英語クエリの生成を行った結果である．“ アバター ”の例では，コンピュータネットワークの用語である“ アバター ”の Wikipedia 記事を参照して英訳が行われた．“ オリックス ”の例では，英訳の際に Wikipedia の「曖昧さ回避のためのページ」が参照された．しかし，企業の“ オリックス ”(ORIX) と動物の“ オリックス ”(Oryx) では英語の綴りが異なり，Wikipedia 上に言語間リンクが存在しない．そのため Wikipedia からは英訳が得られず，Google AJAX Language API から“ Orix ”という英訳が得られた．

表 3 は略称，別称を日本語クエリとして，英語クエリの生成を行った結果である．ここで用いた名称はどれも正式名称ではないため，Wikipedia で検索を行うと正式名称をタイトルに持つ記事にリダイレクトされる．例えば,“ オバマ大統領 ”で検索すると“ バラク・オバマ ”をタイトルとする記事にリダイレクトされる．本例ではそのリダイレクト先で言語間リンクを参照することで，正式名称の英訳を取得した．

表 4 は，物事の方法を調べるときによく使われるクエリについて，英訳を行った結果である．生成された英語クエリの妥当性を確認するため，これらのクエリで実際に検索を行い結果を取得した．一つ目の例では，英語の検索結果の上位 20 件のうち 10 件がアボカドを用いた料理に関するサイトだった．それらにはアボカドを使う料理のレシピの他に，調理法を紹介する動画などが含まれていた．二つ目の例では，英語検索結果中の

表 5　可視化結果の分析に用いるクエリ

検索クエリ (q_1)	検索クエリ (q_2)
“ イチロー ”	“ イチロー　シアトルマリナーズ ”

上位 20 件のうちの 5 件がコーヒーの入れ方に関するサイトであった．他にはコーヒーメーカーの販売サイトなどが含まれていた．

5.1.2　考　　察

表 1 の例のようにクエリ中の固有名詞に対しては，Wikipedia を参照することで適切な英語クエリを生成できることがわかった．Wikipedia の記事中に言語間リンクが存在すれば，辞書では英訳が困難な専門用語や時事語，映画や書籍の題にも対応できる．普通名詞については，Wikipedia に記事が存在しない場合は Google AJAX Language API によって翻訳することになる．“ 支持率 ”は英語サイトでは“ Approval Rating ”という表現が多く見られたが，“ Approval ”を検索語としても検索を行う上で問題はなかった．

表 2 の例のように多義語を翻訳する場合，複数の訳語が候補に挙がることがある．例えば“ オリックス ”という語を翻訳するとき，ユーザが企業の“ オリックス ”を意図して検索していた場合は“ ORIX ”が，動物の“ オリックス ”を意図していた場合は“ Oryx ”が正しい訳語となる．現在の仕様では Google AJAX Language API が返却する結果をそのまま訳語としているが，ユーザの意図を正確に反映するためには訳語をユーザに選択してもらうなどする必要がある．

略称や別称の翻訳については，表 3 のように良い結果が得られた．Wikipedia の記事名は正式名称となっており，略称や別称で検索した場合は正式名称の記事にリダイレクトされる．リダイレクトされた先で言語間リンクを参照することで，略称や別称から正式名称の英訳が得られる．この機能により表記のゆれや類義語，よくある誤表記などを含むクエリにも対応できる．

表 4 の例では，あまり良い英訳といえないものもあった．例えば，クエリの英訳中に出現している“ You ”は有効なクエリとは言えない．また調理法を調べるのなら，“ recipe ”や“ how to ”といった検索語をクエリに追加できれば有効であると考えられる．しかし，これらの検索語は使用した翻訳 API からは得られなかった．

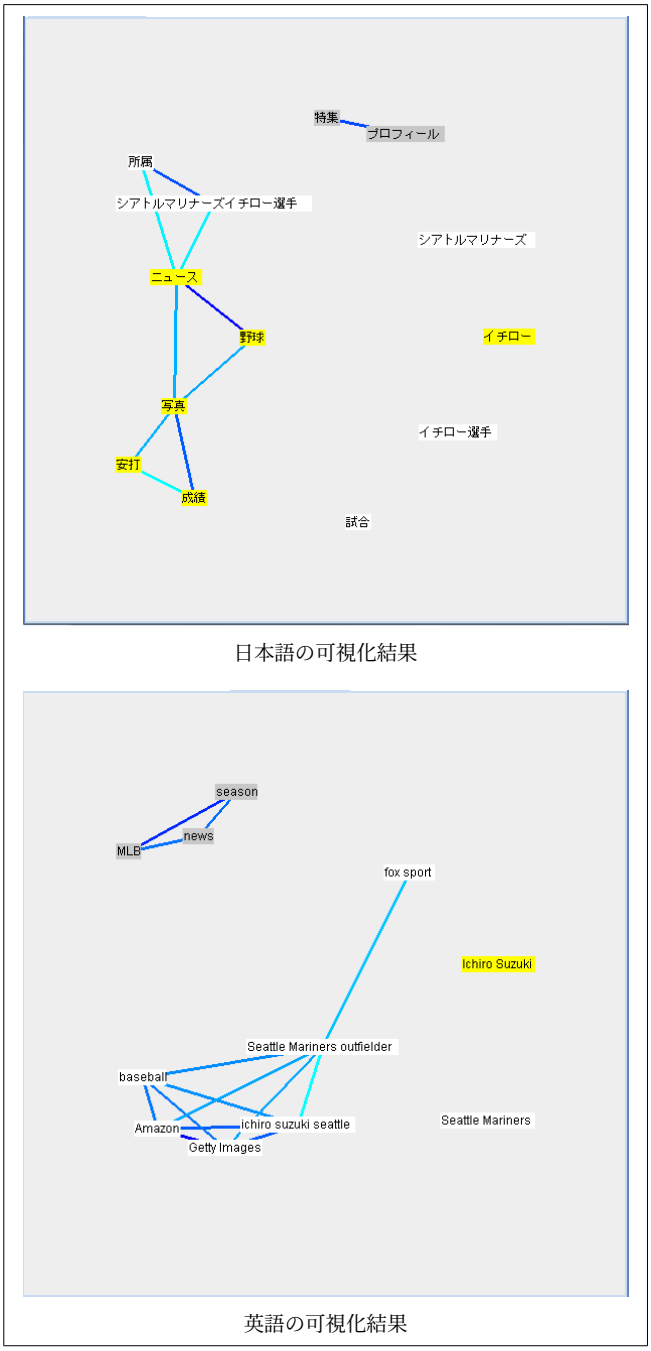
実験結果から，検索語としてよく用いられる時事語や固有名詞に対して本提案手法は有効であると言える．また，別称や略称についても適切な英訳を生成することができる．一方で，多義語についてはユーザの意図に沿う翻訳が行えない場合がある．

5.2　可視化結果の分析

本節では，検索結果の可視化結果について分析する．言語別の可視化と日英二言語を統合した可視化のそれぞれについて，実際にクエリを入力して得られた結果を分析する．検索は 2 回連続して行い，2 回目の検索終了時に提示される可視化結果を分析対象とする．連続する 2 回の検索に用いるクエリ q_1 と q_2 を表 5 に示す．

5.2.1　可視化結果

得られた可視化結果を図 2，図 3，図 4 に示す．図 2 は日英の

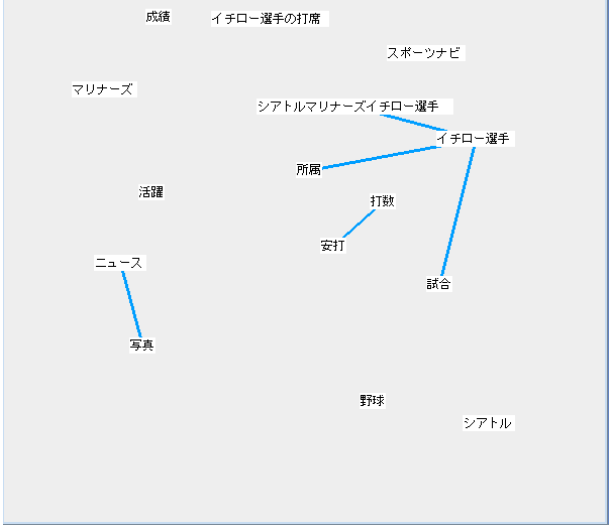


日本語の可視化結果

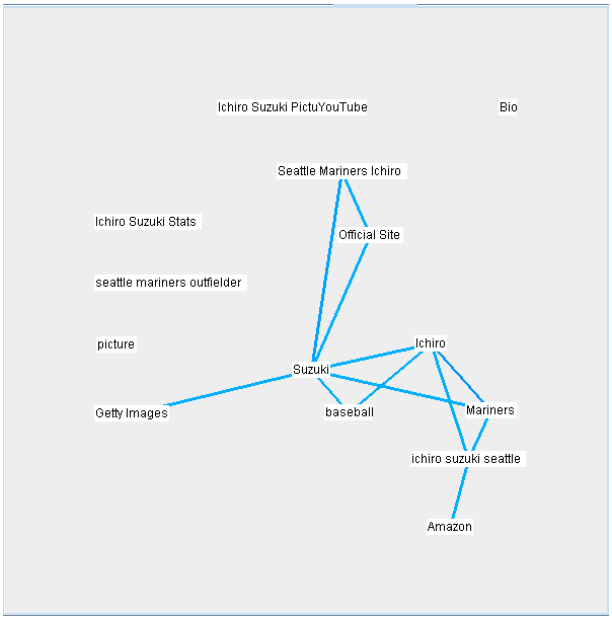
英語の可視化結果

図 2　出現頻度の推移に基づく可視化例

検索結果をそれぞれ個別に，特徴語の出現頻度の推移に基づき可視化した結果である．日本語の可視化結果を見ると，2 回のクエリの両方に出現した“ イチロー ”の他に“ 野球 ”や“ ニュース ”，“ 安打 ”などの特徴語が高い出現頻度を維持していることがわかる．また，2 回目のクエリに追加された“ シアトルマリナーズ ”の他に“ 試合 ”や“ 所属 ”などの特徴語の出現頻度が上昇している．それとは逆に“ 特集 ”と“ プロフィール ”の出現頻度は減少している．高い出現頻度を維持した語の間にはエッジが多く張られており，これは出現頻度の推移が類似していることを示している．この可視化結果から，2 回目の検索ではイチロー選手に関する野球やニュースの話題を維持しつつ，シアトルマリナーズや試合に関する検索結果が増加したことを読み



日本語の可視化結果



英語の可視化結果

図 3　共起度に基づく可視化例

取ることができる．また，ニュース記事に出現する“ 写真 ”や“ 安打 ”,“ 成績 ”といった語は関連性が高いことがわかる．1 回目の検索ではイチロー選手の公式サイト，関連する番組や本，イチロー選手の語録など幅広いページがヒットしていた．それに対して 2 回目の検索では，シアトルマリナーズにおけるイチロー選手に関するページが大半を占めており，他にはシアトルマリナーズの話題を扱うサイトなどがヒットしていた．

英語の可視化結果を見ると，q_1 と q_2 の両方のクエリの訳語に出現した“ Ichiro Suzuki ”は高い出現頻度を維持しており，それ以外の特徴語は全て出現頻度が変化したものである．出現頻度が上昇したものには，新たに q_2 のクエリに含まれた“ Seattle Mariners ”やそれを含む語の他に“ Getty Images ”

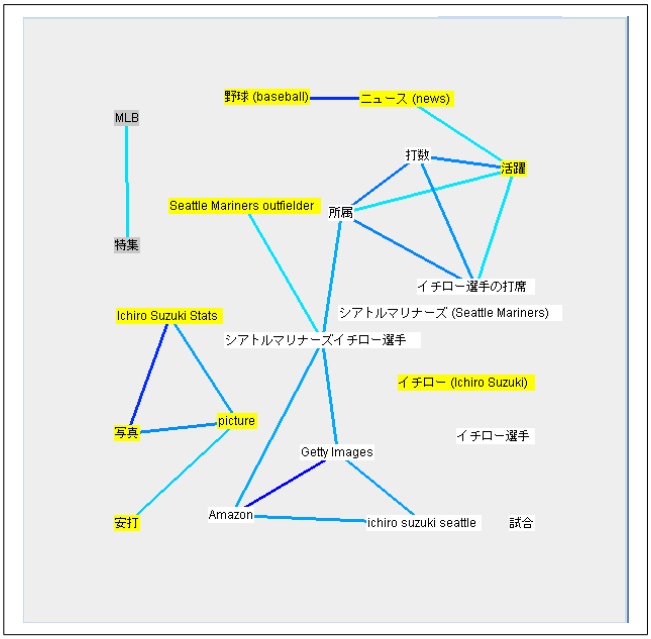


図 4　バイリンガル検索結果の可視化例

や“ Amazon ”などが見られる．逆に出現頻度が減少したものは,“ MLB ”と“ news ”,“ season ”の三つである．エッジは主に出現頻度が増加した特徴語間と出現頻度が減少した特徴語間に見られる．この可視化結果から，2 回目の検索では主に“ seattle Mariners ”に関する話題が増加していることがわかる．“ Amazon ”は通販サイトの Amazon であり，2 回目の検索でイチロー選手の名前が入ったユニフォームの販売ページがヒットしていたために出現頻度が増加したと考えられる．“ Getty Images ”は写真画像代理店であり，試合中のイチロー選手の写真に関するページがヒットしていた．1 回目の検索結果はイチロー選手の公式サイトやファンサイト，MLB における成績のページが多く含まれており，それらのページにはシアトルマリナーズに関係するものも多かった．2 回目の検索でも同じようなページが多くヒットしていたが，よりシアトルマリナーズに関するページが増えており，他にはユニフォームの販売ページなどが新たに出現していた．

図 3 は日英の検索結果をそれぞれ別々に，特徴語の共起度に基づき可視化した結果である．日本語の可視化結果を見ると，共起関係が三つのグループに分かれていることがわかる．一つ目は“ 安打 ”と“ 打数 ”のグループであり，成績を表すときに「5 打数 3 安打」のような表現が多く現れるためだと考えられる．二つ目は“ ニュース ”と“ 写真 ”のグループであり，これはニュース記事には写真が一緒に載せられることが多いためだと考えられる．三つ目は“ イチロー選手 ”と“ 試合 ”,“ 所属 ”,“ シアトルマリナーズイチロー選手 ”のグループである．イチロー選手の話題は彼の所属するシアトルマリナーズやその試合に関連するものが多く，これらの特徴語は共起度が高くなっている．これらのように共起度が高い特徴語は同時に記述されることが多く，関連深い語であることが読み取れる．

英語の可視化結果を見ると,“ Ichiro ”と“ Suzuki ”に多くのエッジが集まっていることがわかる．これらの特徴語はクエリ

中に含まれており，検索結果内で話題の中心になっていると考えられる．他には“ Mariners ”や“ Getty Images ”,“ baseball ”などが表示され，エッジが張られている．

図 4 は日英の検索結果の両方を利用して，特徴語の出現頻度の推移に基づき可視化した結果である．“ イチロー ”と“ Ichiro Suzuki ”,“ シアトルマリナーズ ”と“ seattle Mariners ”など日英間で対訳関係にあるノードは一つに統合されている．これらのように統合されるノードは日英両方の検索結果に出現する語であり，日英共通の話題であると考えられる．逆に“ 打数 ”や“ 安打 ”,“ 活躍 ”といった特徴語は日本語でしか出現していない．このことから，イチロー選手の打数や安打といった記録については日本人の関心の方が強いことがうかがえる．左下部分では四つのノードがエッジで結ばれており，関連があることが読み取れる．“ picture ”と“ 写真 ”は統合できていないものの対訳関係にあり,“ Stats ”(成績) や“ 安打 ”と共にニュース記事に現れるなど，これらの特徴語は関連が深い．エッジが張られている特徴語は出現頻度やその変化が類似しているものであり，言語が異なっているとしてもエッジが張られていれば出現頻度の傾向が似ていると考えられる．

5. 2. 2 考　　察

図 2 の可視化例において，特徴語の出現頻度の推移から検索結果の変化を読み取ることができた．また，出現頻度の推移の類似度をエッジとして可視化することで，変化パターンが類似している特徴語をエッジで結び付けて提示することができた．しかし，エッジがなく他との類似関係が読み取れない特徴語もあり，これらに配慮した可視化を行うことは今後の課題である．

図 3 の可視化例では，特徴語の共起度を用いることでそれらの関連の強さを提示することができた．しかし，こちらの例も図 2 の例と同様にエッジが全くない特徴語がいくつかあり，それらについては他の特徴語との関連性を知ることができなかった．また，同じような単語を含む特徴語が多く提示されており，これらは検索結果の概要を知る上で有効とは言えない．これらの点については今後改善する必要がある．

図 4 の可視化例では，対訳関係を用いて特徴語を統合することにより，日英の検索結果に共通して出現する話題を読み取ることができた．片方の言語にしか出現しない特徴語もあり，その話題についてはその特徴語が出現する言語の検索結果の方が詳しいことがわかる．このようにして日英の検索結果を比較しながら検索結果の概要を知ることができる．

6. ま と め

ユーザが入力した日本語クエリを自動翻訳して日英二言語同時に検索を行い，その結果を可視化する手法を提案した．提案手法では，Wikipedia の言語間リンク等を利用して日本語クエリを自動で英訳した．また，特徴語の共起度と出現頻度の変化に基づき，特徴語をノード，特徴語間の類似関係をエッジとしたグラフを用いて可視化した．そこでは日英言語別の検索結果に基づく可視化と，それらを統合した可視化を行った．

評価実験では生成される英語クエリの妥当性を評価した．その結果，検索クエリ中に頻出する固有名詞や別称，略称に対して正確な英訳が得られることがわかった．また実際の可視化結果について分析し，特徴語の出現頻度の推移を表示することで検索結果の変化を読み取れること，特徴語の共起関係からそれらの語の関連性を読み取れることがわかった．日英二言語の検索結果を両方利用した可視化では，特徴語の対訳関係が日英の検索結果を比較する上で役立つことがわかった．また両言語に出現する特徴語は両言語の検索結果に共通する話題であり，片方の言語にしか出現しない特徴語はその言語の検索結果特有の話題であることも読み取れた．

今後の課題としては，特徴語の抽出規則と特徴語ノードの統合規則を改良してより有効な特徴語を提示すること，ユーザの意図に沿ったクエリの翻訳を行うこと，可視化したバイリンガル検索結果の可読性を向上させる方法について検討することがあげられる．

文　　献

[1] 樋口 重人，福井 雅敏，藤井 敦，石川 徹也，特許情報を対象とした言語横断検索システムの開発，言語処理学会第 7 回年次大会発表論文集，pp.445-447 (2001).

[2] Wikipedia，http://ja.wikipedia.org/

[3] Google AJAX Language API - Google Code, http://code.google.com/intl/ja/apis/ajaxlanguage/

[4] 中崎 寛之，川場 真理子，山崎 小有里，宇津呂 武仁，福原 知宏，同一トピックの日英ブログにおける文化間差異の発見支援，第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009)，A4-4 (2009).

[5] 新井 嘉章，福原 知宏，増田 英孝，中川 裕志，Wikipedia を用いた多言語情報アクセスに関する研究：言語間リンクの分析と応用，第 20 回セマンティックウェブとオントロジー研究会 (2009).

[6] 結城 崇，Web 上のキーワードを辿る視覚的な情報探索インターフェースの開発，筑波大学第三学群情報学類 情報科学主専攻 卒業研究論文 (2008).

[7] 長畑 洋臣，太田 学，検索結果の推移の可視化による検索支援，Web とデータベースに関するフォーラム (WebDB Forum) 2008 論文集，5A-3 (2008).

[8] Yahoo!デベロッパーネットワーク, http://developer.yahoo.co.jp/

[9] Google，http://www.google.co.jp/

[10] sen:ホーム，https://sen.dev.java.net/

[11] 専門用語（キーワード）自動抽出用 Perl モジュール ”TermExtract”の解説, http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

[12] monty tagger :: research :: hugo liu :: a new asthetic, http://web.media.mit.edu/ hugo/montytagger/

[13] Wikipedia:曖昧さ回避，http://ja.wikipedia.org/wiki/%E6%9B%96%E6%98%A7%E3%81