Vol.2010-DBS-151 No.7
2010/11/12

文化間差異理解のためのバイリンガル検索結果の可視化の一手法

内藤　宗一朗[†1]　　太　田　　学[†1]

本研究では，日英二言語によるバイリンガル検索結果を，特に言語間の差異に注目して可視化する手法を提案する．提案手法はユーザが入力した日本語クエリを翻訳して英語クエリを生成し，それら二つのクエリでバイリンガル検索を行う．提案手法はまた，各言語の検索結果から抽出した特徴語間の関係を可視化することで，クエリの表すトピックに対する，日英二言語の文化圏での言及のされ方の違いを容易に把握できるようにする．実験では特徴語の翻訳について，適切な対訳が生成できているか確認した．さらに検索結果の可視化結果について，提示されるノードとエッジの数を調査した．

A Visualization Method of Bilingual Search Results for Understanding Cultural Differences

Soichiro Naito[†1] and Manabu Ohta[†1]

We propose a method to visualize bilingual search results in Japanese and English with focusing on the differences between them. The proposed method automatically generates an English query by translating the Japanese query inputted by a user and searches the Web using both queries. In addition, the method visualizes relations among the feature terms extracted from the search results. We evaluated the quality of translation of the extracted feature terms. We also evaluated visualized search results by investigating the number of visualized nodes and edges.

†1 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

1. はじめに

あるトピックに関する情報を入手する手段の一つとして，サーチエンジンが広く利用されている．Google[⋆1]や Yahoo!検索[⋆2]をはじめとする主要なサーチエンジンには，検索対象とするウェブサイトの言語を指定する機能を持つものも存在する．この機能を用いることで特定のトピックに関する情報を言語別に収集することが可能であるが，その検索結果の内容は言語毎に大きく異なることも珍しくない．この差異は，あるトピックに対する言語文化間の言及の差異といえる．

そこで本研究では，日英二言語によるバイリンガル検索の結果を，言語間の差異に着目して可視化する手法を提案する．提案手法では，ユーザがトピックとして入力した日本語クエリを翻訳して英語クエリを生成し，日英二つのクエリでバイリンガル検索を行う．それにより得られた日英それぞれの検索結果から特徴語を抽出し，その特徴語間の関係を可視化することで，トピックに関する言語文化間差異の理解を支援する．可視化は特徴語をノード，特徴語の共起関係をエッジとするグラフにより行う．ノードはバネモデルにより配置し，また手動でのノードの再配置も可能にする．また，英語に精通していない利用者が手軽に利用できるように，クエリの入力から可視化結果の閲覧まで可能な限り日本語で行えるのは提案手法の主要な特長の一つである．このようにバイリンガル検索結果の差異と共通点を可視化できれば，例えばある商品についての国内と海外での評判の差を容易に把握することで，マーケティングリサーチに応用できる

本稿では，まず 2 節で関連研究について述べる．3 節で提案手法，4 節で評価実験について述べ，最後に 5 節でまとめと今後の展望について述べる．

2. 関連研究

まず，本研究に関連する研究について述べる．2.1 節では言語横断検索に関する研究，2.2 節では検索結果の可視化に関する研究について説明する．

2.1 言語横断検索

対訳辞書を利用してクエリを翻訳し，特許データベースの検索を行う研究[1)] がある．この研究では，あらかじめ用意した対訳辞書を用いてクエリの翻訳を行っている．また，対訳

⋆1 http://www.google.co.jp/
⋆2 http://search.yahoo.co.jp/

辞書から抽出した翻訳モデルと特許データベースから抽出した言語モデルを用い，訳語の曖昧性を解消する方法を提案している．さらに日米対応特許を多言語コーパスとして利用し，対訳辞書を更新する機能を実装している．

また，多言語情報へのアクセス支援を目的とした多言語対訳システムの研究[2] もある．この研究では，Wikipedia[⋆1]の言語間リンクを利用して多言語対訳システムを構築している．また，システムから得られる訳語の数や質を調査するために，Wikipedia の言語間リンクの詳細な分析を行い，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．

ある同一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研究[3] もある．この研究では Wikipedia の言語間リンクを対訳表現の取得に利用し，検索を行っている．さらに検索結果から共起語を抽出し，それらを出現確率と言語間の出現比率に基づきマッピングして可視化している．

本研究でも日本語クエリの英訳，英語特徴語の和訳には Wikipedia の言語間リンクを利用する．また検索対象は Web サイト全般とし，普段使われることの多い一般的な Web 検索に対して提案手法の適用を試みる．さらに本研究では取得した日英二言語の検索結果について，特徴語と特徴語間の関係をグラフを利用して可視化して，ユーザに提示する．

2.2 検索結果の可視化

ベン図を用いて検索結果を可視化する研究[4] がある．この研究では，入力されたクエリの関連語をベン図上に表示される円に対応させ，各関連語をクエリとした際の検索ヒット数を円の大きさで表現している．また円同士の重なり合いの大小により，各クエリによる AND，OR，NOT 検索のヒット数を表現している．

本研究では，クエリの関連語ではなく検索結果から抽出される特徴語を可視化して，その結果をユーザに提示する．また特徴語をノード，特徴語間の関係をエッジとするグラフを用いて検索結果を可視化しているため，ユーザは特徴語間の関係が把握し易く，またノードを再配置することもできる．

長畑らの研究[5] では，検索結果から抽出される特徴語と特徴語間の関係を可視化している．彼らは，連続した検索における特徴語を出現頻度の推移に基づき三つに分類し，その推移のパターンの類似度を可視化することで，ユーザの検索支援を行った．結果は特徴語をノード，推移の類似度をエッジとするグラフの形で可視化され，ノードの配置にバネモデルを利用することで結果の可読性の向上を試みている．

⋆1 http://ja.wikipedia.org/

本研究では日英二言語の検索結果から特徴語を抽出し，特徴語の共起関係と言語間の差異に着目した可視化を行う．日英二言語を対象に検索を行う点が長畑らの研究との主な相違点である．また本研究では，特徴語の出現頻度の推移ではなく，日英二言語間の検索結果の差異に着目して可視化する．

3. 提 案 手 法

提案手法は，ユーザが入力した日本語クエリから英語クエリを自動生成し，ユーザに英語クエリを意識させることなく日英二言語で検索する．このような検索を本稿ではバイリンガル検索と呼んでいる．提案手法はまた，バイリンガル検索により得られた検索結果を可視化してユーザに分かり易く提示する．可視化はグラフを用いて行い，検索結果から抽出した特徴語をノード，特徴語の共起関係をエッジとする．また，日英の検索結果の差異を明示することで，クエリに対する日英間での言及のされ方の違いを，ユーザに理解してもらうことを目的とする．二言語間の検索結果の共通項と差異は，それぞれの検索結果から抽出された特徴語について対訳関係を調査することで取得する．

以下に提案手法の大まかな流れを示す．

(1) ユーザの入力した日本語クエリを取得
(2) 取得した日本語クエリを英訳して英語クエリを生成
(3) 日英二言語で検索
(4) 得られた日英の検索結果から特徴語を抽出
(5) 日英の特徴語間で対訳関係を調査
(6) 特徴語の共起関係と対訳関係を基に検索結果を可視化

3.1 バイリンガル検索

本節では，ユーザが入力した日本語クエリから英語クエリを生成し，日英二言語で検索を行う手法[6] について述べる．提案手法では，入力された日本語クエリを英訳することで英語クエリを生成する．英訳には Wikipedia の言語間リンクと Google AJAX Language API[⋆2]を利用する．これらを利用して生成された英語クエリと元の日本語クエリを用い，検索を行う．

3.1.1 英語クエリの生成

提案手法は入力として日本語のクエリを想定している．入力された日本語クエリをスペー

⋆2 http://code.google.com/intl/ja/apis/ajaxlanguage/

スを区切りとして検索語単位に分割し，それらを英訳することで英語クエリを生成する．検索語の英訳は Wikipedia の言語間リンクを利用して行う．Wikipedia の言語間リンクとは，異なる言語で書かれた同じ主題の記事同士を結び付けるリンクのことである．このとき言語間リンクで結ばれた各記事のタイトルは対訳関係にある．提案手法では，言語間リンクで結ばれた日英 Wikipedia 記事のタイトルを対訳辞書として用い，クエリの英訳に利用する．Wikiepdia には多くの専門用語や固有名詞，最新の事柄についての記事が存在する．英訳に Wikipedia の言語間リンクを利用することで，辞書翻訳が難しいこれらの語も英訳することができる．

また，Wikipedia にはリダイレクトと呼ばれる機能が存在する．通常，Wikipedia の記事タイトルは正式名称である．Wikipedia におけるリダイレクトとは，ある物の略称や別称で記事を検索した際に正式名称の記事に転送する機能である．提案手法では，このリダイレクト機能を言語間リンクと合わせて利用することで，略称や別称がクエリ内に含まれていても正式名称の英訳を生成することができる．

提案手法ではまた，クエリの英訳に Google AJAX Language API を利用する．Google AJAX Language API は，テキストを指定した言語に翻訳する機能を備えており，この機能をクエリの英訳に利用する．Wikipedia の言語間リンクで対訳が取得できない場合に，Google AJAX Language API を利用することで，英語クエリを生成することができる．

3.1.2 日英二言語のクエリによる検索

ユーザが入力した日本語クエリと提案手法で生成した英語クエリの二つを用いて，日英二言語で検索を行う．検索には Yahoo!ウェブ検索 Web API[⋆1]を利用する．Yahoo!ウェブ検索 Web API を利用する際に対象とする言語を指定することで，日本語クエリからは日本語の，英語クエリからは英語の検索結果を取得することが出来る．

3.2 特徴語抽出

本節では，バイリンガル検索により取得した検索結果から特徴語を抽出する方法[6] について述べる．3.2.1 項では日本語検索結果から特徴語を抽出する方法について，3.2.2 項では英語検索結果から特徴語を抽出する方法について説明する．

3.2.1 日本語検索結果からの特徴語抽出

日本語検索結果からの特徴語抽出は長畑らの方法[5] を参考にした．まず形態素解析器 Sen[⋆2]を用いて，検索結果のタイトルとサマリから形態素を抽出する．得られた形態素の中から，以下のものを抽出して連結することで特徴語を生成する．

(1) 名詞
(2) カタカナのみで構成される未知語
(3) 英数字のみで構成される形態素
(4) 接頭辞
(5) 区切り記号「・」,「-」,「.」
(6) 連体助詞「の」

生成される特徴語は，日本語の特徴語と外国語の特徴語に分けられる．(1)，(2)，(4)，(5)，(6) を連結する場合は日本語の特徴語とする．(3)，(5) を連結する場合は外国語の特徴語とする．生成される特徴語が長くなりすぎることを防ぐため，接頭辞の前及び接尾辞の後ろには連結しない．

3.2.2 英語検索結果からの特徴語抽出

英語検索結果からの特徴語抽出は TermExtract[⋆3] を参考にした．まず Monty Tagger[⋆4]を用いて検索結果のタイトルとサマリを単語に分割し，それらに品詞タグを付与する．品詞タグが付与された単語のうち，以下のものを特徴語の構成要素とする．

(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) 動詞の過去分詞形

これらの単語を連結することで特徴語を生成する．生成される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．

3.2.3 特徴語のスコア付け

3.2.1 項，3.2.2 項で述べた方法で抽出した特徴語にスコア付けを行い，各言語中のスコアの上位 10 語ずつを可視化対象の特徴語とする．スコアは TF 及び TF-IDF により算出する．

⋆1 http://developer.yahoo.co.jp/
⋆2 https://sen.dev.java.net/
⋆3 http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
⋆4 http://web.media.mit.edu/~hugo/montytagger/

また可視化対象となった特徴語については，それぞれの言語の検索結果内の特徴語同士で共起度を算出する．特徴語 w_i と w_j の共起度 $coo_{i,j}$ を式 (1) により定義する．ここで N は取得した検索結果の件数である．$co\_occur^d_{i,j}$ は，式 (2) に示すように d 番目の検索結果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．

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

$$co\_occur^d_{i,j} = \begin{cases} 1 & \text{if } w_i \in D_d \ \wedge w_j \ \in D_d \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

3.3 検索結果の可視化

本節では，抽出した特徴語を基に検索結果を可視化する方法について述べる．提案手法では，抽出した日英二言語の特徴語を対訳関係を基に分類する．さらに，特徴語間の共起関係を用いることで，グラフによる検索結果の可視化を行う．

3.3.1 対訳関係に基づく特徴語の分類

検索結果を日英二言語間で比較するために，対訳関係に基づき抽出した特徴語の分類を行う．提案手法では，各言語の検索結果から抽出される特徴語の違いによって言語間の検索結果の差異を表す．そこで 3.2.3 項で可視化対象とした日英 10 語ずつの特徴語を，日本語の検索結果のみで出現する語，英語の検索結果のみで出現する語，日英両方の検索結果で出現する語の 3 種類に分類する．日英の特徴語間で対訳関係にあるものを探し，対訳関係が見つかった特徴語は日英両方の検索結果に出現する語となる．対訳関係が見つからなかった特徴語は日本語もしくは英語の検索結果のみで出現する語となる．対訳関係は，英語特徴語の和訳を日本語特徴語と比較することで発見する．英語特徴語を和訳する方法は 3.3.2 項で説明する．

3.3.2 英語特徴語の和訳

英語特徴語の和訳は，3.1.1 項で説明した日本語クエリの英訳と同様に Wikipedia の言語間リンクを用いて行う．Wikipedia の言語間リンクは双方向リンクであることが多いので，英訳だけでなく和訳にも利用することができる．この方法により得られる英語特徴語の和訳を，3.3.3 項で述べる英語特徴語の可視化で利用する．また，日本語クエリの英訳では Wikipedia の言語間リンクと共に Google AJAX Language API も利用するが，英語特徴語の和訳では Wikipedia の言語間リンクのみを用いる．

3.3.3 共起関係と対訳関係に基づく可視化

検索結果の可視化はグラフを用いて行う．抽出した特徴語をノード，特徴語間の共起度の強さをエッジで表す．ノードの配置はバネモデル[5] によって行い，可読性の向上を図る．3.3.1 項で述べたように特徴語は 3 種類に分類されるので，その分類に応じてノードを 3 色で塗り分ける．日英両方の検索結果に共通する特徴語は，対訳関係にある特徴語を一つのノードに統合して表示する．また，英語検索結果から抽出された特徴語は，3.3.2 項の方法で和訳されたものがノードとして表示される．複数の英語特徴語に対して同じ和訳が得られた場合，それらを一つのノードに統合する．ただし，Wikipedia の言語間リンクで和訳が取得できなかった特徴語については，英語のままノードとして表示する．これにより可視化結果の多くを日本語で閲覧することが可能になる．

図 1 は実装したプロトタイプシステムの実行画面である．クエリ「マチュピチュ」の検索結果と，その可視化結果が表示されている．実行画面の左上部分には，クエリ入力欄や検索ボタンなどの操作系統が存在する．左下には可視化結果の表示領域が，右には検索結果の表示領域が配置されている．可視化結果については，日英の特徴語を別々に可視化したものと，日英の特徴語を統合して可視化したものを選択することができる．日英の特徴語を統合した可視化では，ノードの色分けによって特徴語の分類を表す．青色のノードは日本語検索結果のみに出現した特徴語であり，桃色のノードは英語検索結果のみに出現した特徴語である．そして黄色のノードは日英両方の検索結果に出現した特徴語である．特徴語ノードはバネモデルによる自動配置に加え，ユーザの手による配置変更も可能である．

4. 評 価 実 験

本節では，可視化結果の分析と評価，翻訳精度の評価について述べる．可視化結果については，検索結果と見比べて可視化されている内容を分析し，またノードやエッジの数について調査する．翻訳精度の評価では，英語特徴語の和訳精度を評価する．

4.1 可視化結果の分析

具体的なクエリを用いてバイリンガル検索を行い，その可視化結果について分析する．4.1.1 項ではクエリ「スポーツ」，4.1.2 項ではクエリ「賭博」の可視化結果について述べる．

4.1.1 クエリ「スポーツ」の可視化結果

クエリとして「スポーツ」を入力し，得られた可視化結果を図 2 に示す．これは，2010 年 10 月 6 日時点の検索結果を可視化したものである．

日本語の検索結果には，スポーツニュースを扱うサイトが多く含まれている．日本語特

Vol.2010-DBS-151 No.7
2010/11/12

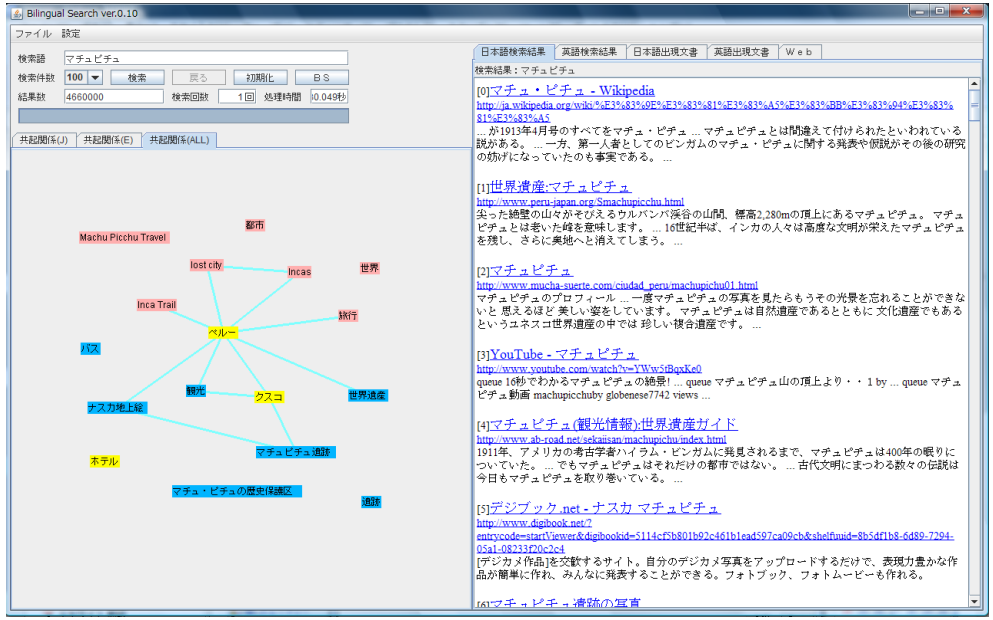

図 1 実行画面
Fig. 1 Screenshot of the prototype system.

有の特徴語としては「野球」や「サッカー」,「格闘技」,「モータースポーツ」などが抽出されており，これらはスポーツニュースサイトで主に扱われる競技である．また「コラム」や「スポーツ総合」といった語も抽出されているが，これらはスポーツニュースサイトのメニュー項目に含まれている語である．

英語の検索結果としても日本語と同様に，多くのスポーツニュースサイトがヒットした．しかし特徴語として抽出された競技名は日本語のものとは異なり,「テニス」や「クリケット」,「ラグビー」,「フットボール」が抽出されている．英語のスポーツニュースサイトでは，主にこれらの競技が扱われていることがわかる．他には「ビデオ」や「結果」といった特徴語が表示されており，これらもスポーツニュースサイトの中で出現する語である．

日英に共通する特徴語としては,「ゴルフ」と「ニュース」があった.「ニュース」は日英両方の言語でニュースサイトが多くヒットしているため，特徴語として抽出されている.「ゴルフ」は日英両方のスポーツニュースサイトで扱われる競技であるため，共通する語として

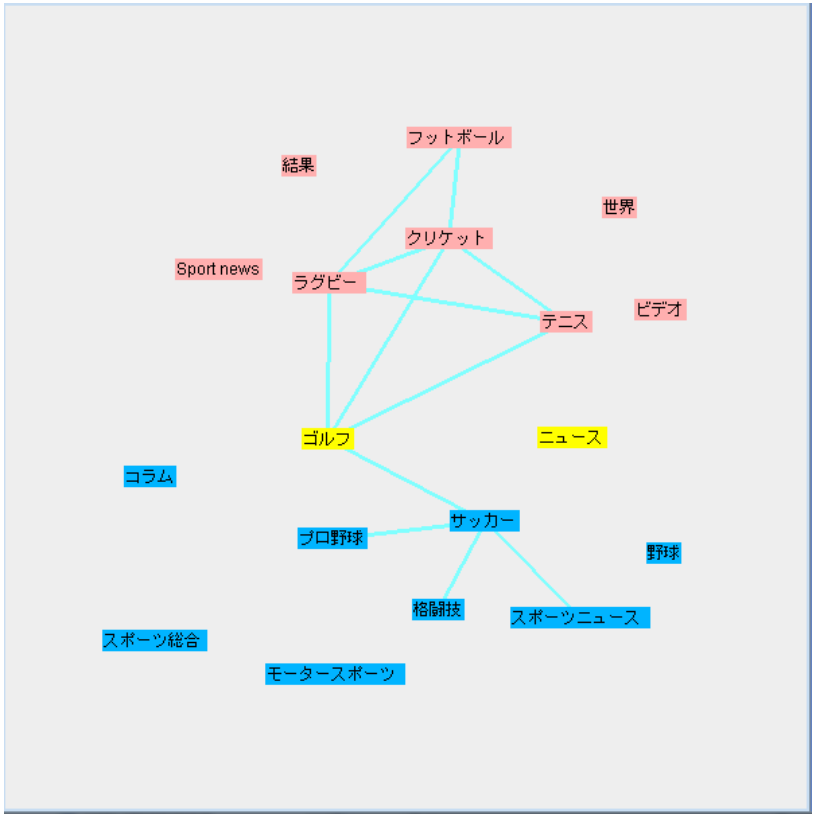

図 2 クエリ「スポーツ」の可視化結果の例
Fig. 2 An example of visualization results of a query “ sport ”.

表示されている.「スポーツニュース」と「Sport news」も日英に共通する語であるが，対訳関係を取得できなかったため，日英それぞれの特徴語として表示されている．一方,「サッカー」と「フットボール」にも対訳関係はあるが，同じ英語圏でも英国と米国では「フットボール」が指す競技は異なっており，一概に「サッカー」と「フットボール」を同一視することはできない．

この「スポーツ」のバイリンガル検索結果の可視化結果から，日英の検索結果には共通してスポーツニュースを扱うサイトが多く含まれていることがわかる．また，それらのスポー

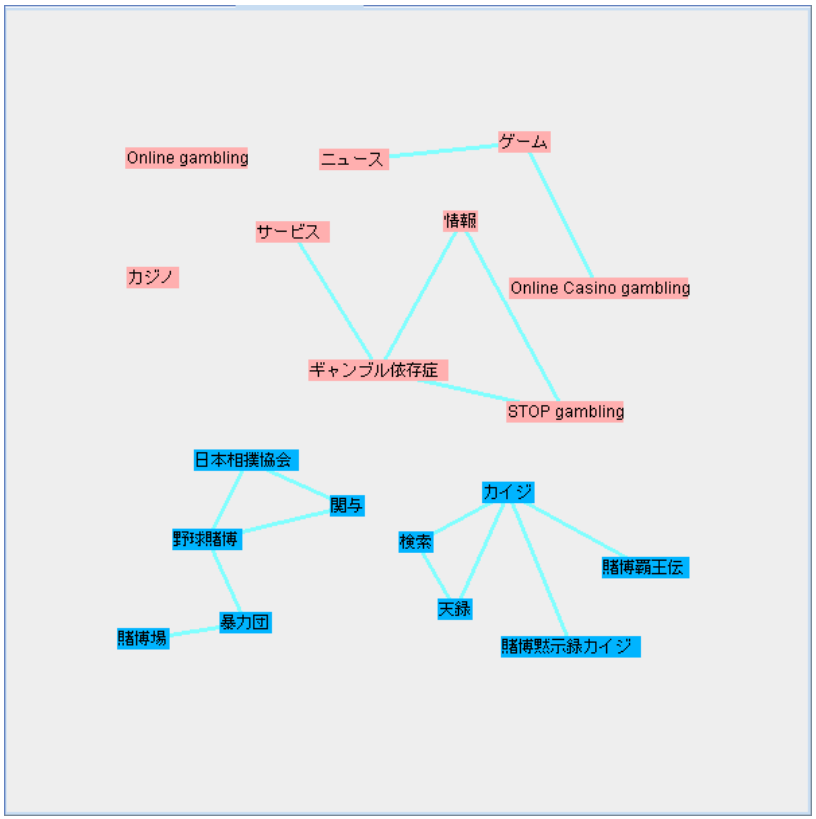

図 3 クエリ「賭博」の可視化結果の例
Fig. 3 An example of visualization results of a query “ gambling ”.

ツニュースサイトが主に扱う競技は，日英で異なることもわかる．

4.1.2 クエリ「賭博」の可視化結果

クエリ「賭博」の可視化結果を図 3 に示す．これは，2010 年 8 月 27 日時点の検索結果を可視化したものである．本例では，日本語と英語の検索結果に共通する特徴語が提示されなかった．「カジノ」と「賭博場」は近い意味を持っているが，提案手法では対訳関係にあるとはされていない．

日本語特有の特徴語としては，10 語の特徴語が可視化されている．共起度を表すエッジを見ると，これらの特徴語は二つのグループに分けることができる．「野球賭博」や「日本相撲協会」，「暴力団」等の特徴語が含まれるグループと，「賭博黙示録カイジ」や「賭博覇王伝」等の特徴語が含まれるグループがある．前者のグループの特徴語は，主に相撲界の野球賭博問題に関する話題の中で出現している．後者のグループの特徴語は賭博を題材とした漫画に関する語であり，前者のグループとは異なった話題の中で出現している．

一方，英語特有の特徴語中には，「Online gambling」や「Online Casino gambling」といった語が見られる．日本ではオンライン賭博やオンラインカジノに関する話を耳にすることは少ないが，英語圏ではオンライン賭博やカジノはより身近なもののようである．「Online gambling」という記事が英語 Wikipedia には存在し，日本語 Wikipedia には対応する記事が存在しないことからも，その普遍性の違いが伺える．他には「ギャンブル依存症」や「STOP gambling」といった特徴語も見られ，そのような話題が検索結果中に多く含まれていることがわかる．また，英語の特徴語が 9 語しか表示されていないが，これは「gambling problem」と「problem gambling」という二つの特徴語が抽出され，それらに同じ「ギャンブル依存症」という和訳があてられ，一つのノードに統合されたためである．

この「賭博」のバイリンガル検索の可視化結果から，日本語の検索結果には相撲界の野球賭博問題や賭博漫画に関する話題が多く含まれているのに対し，英語の検索結果ではオンライン賭博やギャンブル依存症について多く言及されていることがわかる．

4.2 翻訳精度の評価

Wikipedia の言語間リンクと Google AJAX Language API を用いた日本語クエリの英訳については，一定の翻訳精度が得られることを実験により示した[6]．そこで本節では，Wikipedia の言語間リンクを用いた，英語特徴語の和訳の精度について評価する．評価対象として，4.1.1 項と 4.1.2 項の実験において抽出された英語特徴語と，その和訳結果を用いる．評価対象となる英語特徴語は合計 20 語である．

正確な和訳が得られた英語特徴語は，20 語のうち 16 語であった．これらの語は英語検索結果の中における意味と和訳が一致しており，英語特徴語を日本語で正確に表すことができている．一方，正確な和訳が得られなかった特徴語は，表 1 にまとめた 4 語である．いずれも訳語を誤ったのではなく，そもそも訳語が得られていない．このうち「Online gambling」と「Sport news」については，英語の Wikipedia には該当する記事が存在するものの，日本語の Wikipedia に該当する記事は存在しない．そのため言語間リンクが張られておらず，和訳を得ることができなかった．「Online Casino gambling」と「STOP gambling」は英語の Wikipedia 中にも該当記事が存在せず，和訳を得ることができなかった．

Vol.2010-DBS-151 No.7
2010/11/12

表 1 和訳できなかった英語特徴語
Table 1 The English feature terms the prototype system failed in translating into Japanese.

英語特徴語	和訳結果
Online gambling	(和訳無し)
Online Casino gambling	(和訳無し)
STOP gambling	(和訳無し)
Sport news	(和訳無し)

表 2 可視化結果の統計情報
Table 2 Statistics of visualization results.

全ノード数	18.7
全エッジ数	11.1
対訳関係による統合ノード数（黄色のノード数）	1
ノードの次数	1.19

この結果から，Wikipedia の言語間リンクを利用した英語特徴語の和訳については，一定の有効性が確認できる．しかし，先ほど挙げたような和訳が困難な複合語や，多義語の和訳にはまだ問題がある．複合語については，単語単位に区切って翻訳する方法や他の翻訳ツールの適用を検討している．多義語については，共起語を利用して最もふさわしい訳語を得る方法を考えている．また和訳の対象となる英語特徴語は，なるべくノイズを含まない適切な形で抽出されていることが望ましいため，特徴語抽出手法の改善も検討している．

4.3 可視化結果の評価

プロトタイプシステムを利用して，可視化結果として得られるグラフについて全ノードの数，全エッジの数，対訳関係による統合ノードの数，ノードの次数を調査した．本実験には,「スポーツ」と「賭博」に「イチロー」,「イルカ」,「カジノ」,「クジラ」,「ノーベル賞」,「ラーメン」,「リーマンショック」,「ワールドカップ」を加えた 10 例のクエリを用い，それらの結果の平均を算出した．

結果は表 2 のようになった．表 2 を見ると，可視化されている特徴語ノードの中で，対訳関係が見つかり統合できたものは少ない．このことから，対訳関係にあるクエリで検索を行っても，日英言語間で得られる検索結果にかなり差異があることがわかる．よって提案手法のように，言語間の検索結果の差異を抽出するシステムには意義があると考える．しかし，近い意味を持ちながら統合できない,「賭博場」と「カジノ」などの特徴語も存在する．このような特徴語についても，対訳関係を取得して統合できるよう手法を改善したい．

また，ノードの次数の平均は 1.19 となっている．よって，平均すると各ノードの次数は 1 以上ということになる．しかし実際の可視化結果では，ノードの次数に偏りがある例も見られた．次数が 0 のノードが，可視化結果中に 11 個存在している例もあった．これは特徴語を可視化する際に，出現頻度が高い特徴語から順に可視化し，特徴語間の共起度を考慮していないことが理由であると考えられる．他の特徴語と頻繁に共起する語は重要な語であると考えられるため，可視化する特徴語を選択する際には，特徴語の出現頻度だけでなく他の語との共起度も考慮する必要がある．

本稿のプロトタイプシステムでは，言語毎に特徴語を 10 語ずつ選択し，その中で対訳関係や共起関係を取得する．しかし一度の検索で抽出される特徴語は通常 100 語以上であり，その大半は対訳関係や共起関係が考慮されていない．よって今後は，可視化されていない特徴語についても出現頻度や共起度を数値化して評価し，可視化結果の可読性も踏まえつつ提示する特徴語数を検討する必要があると考える．

5. ま と め

本稿では，ユーザが入力した日本語クエリを基に日英二言語でバイリンガル検索を行い，両言語の検索結果の差異に着目して検索結果を可視化する手法を提案した．Wikipedia の言語間リンクを対訳辞書として利用することで，日本語クエリの英訳や日英特徴語間の対訳関係の同定を行った．

評価実験では,「スポーツ」と「賭博」というクエリを用いてバイリンガル検索を行い，得られた可視化結果を分析した．可視化された特徴語は検索結果中の主要な語であり，日英二言語の特徴語を比較することで，両言語の検索結果の差異の把握に役立つことがわかった．可視化結果に表示されるグラフのノード数やエッジ数についても，実験を行って調査した．可視化結果には，日本語もしくは英語の検索結果だけから抽出される特徴語が多数提示されており，対訳関係にあるクエリを用いて検索しても，日本語と英語では検索結果に大きな差が見られることがわかった．また，Wikipedia の言語間リンクを用いた英語特徴語の和訳精度も評価した．20 語の英語特徴語のうち 16 語で適切な和訳が得られ，Wikipedia の言語間リンクによる和訳の一定の有効性を確認した．

今後の課題としては，提示する特徴語の数や選択方法の改善が挙げられる．特徴語のスコア付けと抽出規則を改善することで，より適切な特徴語の提示が可能になると考えられる．スコア付けは出現頻度だけでなく，他の語との共起度や対訳関係も考慮して行うことで，より有効な特徴語が選定できるのではないかと考える．英語特徴語の和訳については，複合語

Vol.2010-DBS-151 No.7
2010/11/12

や多義語への対応など改善の余地がある．また，現状の可視化結果はノードを言語別に色分けしただけの簡素なものなので，さらに可読性を向上させるための可視化方法についても検討したい．例えば，特徴語だけをユーザに提示するのではなく，その詳細情報や関連情報を特徴語に付与できれば，内容の把握がより容易になると考えている．

参 考 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) 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， SIG-SWO-A803-15 (2009).
3) 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，DEIM フォーラム論文集， A4-4 (2009).
4) 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 平成 20 年度卒業論文 (2008).
5) 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).
6) 内藤宗一朗，太田　学：可視化によるバイリンガル検索支援，DEIM フォーラム論文集， D10-1 (2010).