Wikipediaエントリを知識源とする 日英ブログからの文化間差異発見支援

Semi-Automatic Discovery of Cross-Cultural Gaps from Japanese/English Blogs with Wikipedia as Fundamental Knowledge Source

中崎 寛之[*1] 川場 真理子[*2] 宇津呂 武仁[*1] 福原 知宏[*3]
Hiroyuki Nakasaki Mariko Kawaba Takehito Utsuro Tomohiro Fukuhara

[*1]筑波大学大学院システム情報工学研究科
Grad. Sch. Systems and Information Engineering, University of Tsukuba

[*2]日本電信電話株式会社 NTT サイバースペース研究所
NTT Cyber Space Laboratories, NTT Corporation

[*3]東京大学 人工物工学研究センター
Research into Artifacts, Center for Engineering, University of Tokyo

The goal of this paper is to cross-lingually analyze multilingual blogs collected with a topic keyword. The framework of collecting multilingual blogs with a topic keyword is designed as the blog feed retrieval procedure. Multilingual queries for retrieving blog feeds are created from *Wikipedia* entries. Finally, we cross-lingually and cross-culturally compare less well known facts and opinions that are closely related to a given topic. Preliminary evaluation results support the effectiveness of the proposed framework.

1. はじめに

本論文では，ある同一のトピックについてまとまった規模の記述が書かれたブログサイトを，日英各言語について検索し，その記述内容を二言語間で対照分析する方式を提案する (図 1) [中崎 09]．あるトピックの日英二言語表現を得る際には，Wikipedia の日英二言語エントリを用いる．ブログサイトの検索においては，特定トピックを表すキーワードを用いて商用検索エンジン API により上位のブログサイトを収集し，これを，特定トピックを表すキーワード，および Wikipedia から収集した関連語の出現数順にリランキングする方法を用いる [川場 08, 川場 09]．この方法により，そのトピックについての記述が多く含まれる有用なブログサイト，および，それらのブログサイト中における有用な記事を上位にランキングすることが可能となる．さらに，これまでに行った評価実験では，それらのブログサイトの内容を日英二言語間で対照分析することにより，ブログ特有の個人レベルの情報や意見における国間差異が多数観測されている．

2. 評価用トピック

評価用トピック候補として，Wikipedia において，日英 Wikipedia エントリが存在し，日英ブログ空間におけるエントリ名のヒット数が一定数以上となるものを 50 個程度選定した．このトピック候補の中から，評価用トピックとして，「捕鯨」，「臓器移植」，「喫煙」，「サブプライムローン」の社会系トピック 4 種類を選定した．これらの評価用トピックの要約と日英ブログにおける評価用トピックに対する主な意見を表 1 に示す．

連絡先: 中崎 寛之，筑波大学大学院システム情報工学研究科，〒 305-8573 茨城県つくば市天王台 1-1-1，029-853-5427

3. 二言語対照ブログ分析

3.1 ブログサイト検索

Wikipedia エントリをトピックとするブログサイトの検索においては，日本語ブログの検索には，Yahoo!Japan 検索 API を，英語ブログの検索には米 Yahoo!検索 API を利用し，日本語ブログでは大手 11 社[*1]，英語ブログでは大手 12 社[*2]のブログ会社のドメインに限って検索を行った．検索の際には，Wikipedia エントリのエントリ名を検索クエリとして，複数のブログホストを一度に指定して検索し，1000 件の記事を取得する．しかし API の検索ではブログ記事単位の検索になるので，同一著者のブログ記事は一つのブログサイトにまとめるという作業を行った．その結果，一トピックあたり約 200 前後のブログサイトを取得することができた．その後，各ブログサイトにおいて，Wikipedia エントリのエントリ名のヒット数を求め，ヒット数が下限未満 (本論文では，10) のブログサイトを削除した．

3.2 ブログ記事検索

次に，検索した日英ブログサイト集合の中から，トピックについて詳しく書かれたブログ記事を検索する．手法としては，トピック名がタイトルである各言語の Wikipedia エントリのリダイレクト，さらに Wikipedia エントリの本文から太字，他エントリリンクをブログ記事検索のための関連語として抽出する．そして，抽出した関連語のいずれかが出現する各言語のブログ記事をブログサイト集合内からそれぞれ検索する．各トピックの Wikipedia 関連語数，各トピックで検索したブログサイト数，検索したブログサイト中で Wikipedia 関連語のいずれかが出現したブログ記事数，検索したブログ記事本文に含まれる総形態素数および総単語数を表 2 に示す．

∗1 FC2.com, yahoo.co.jp, rakuten.ne.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, jugem.jp, yaplog.jp, webry.info.jp, hatena.ne.jp

∗2 blogspot.com, msnblogs.net, spaces.live.com, livejournal.com, vox.com, multiply.com, typepad.com, aol.com, blogsome.com, wordpress.com, blog-king.net, blogster.com

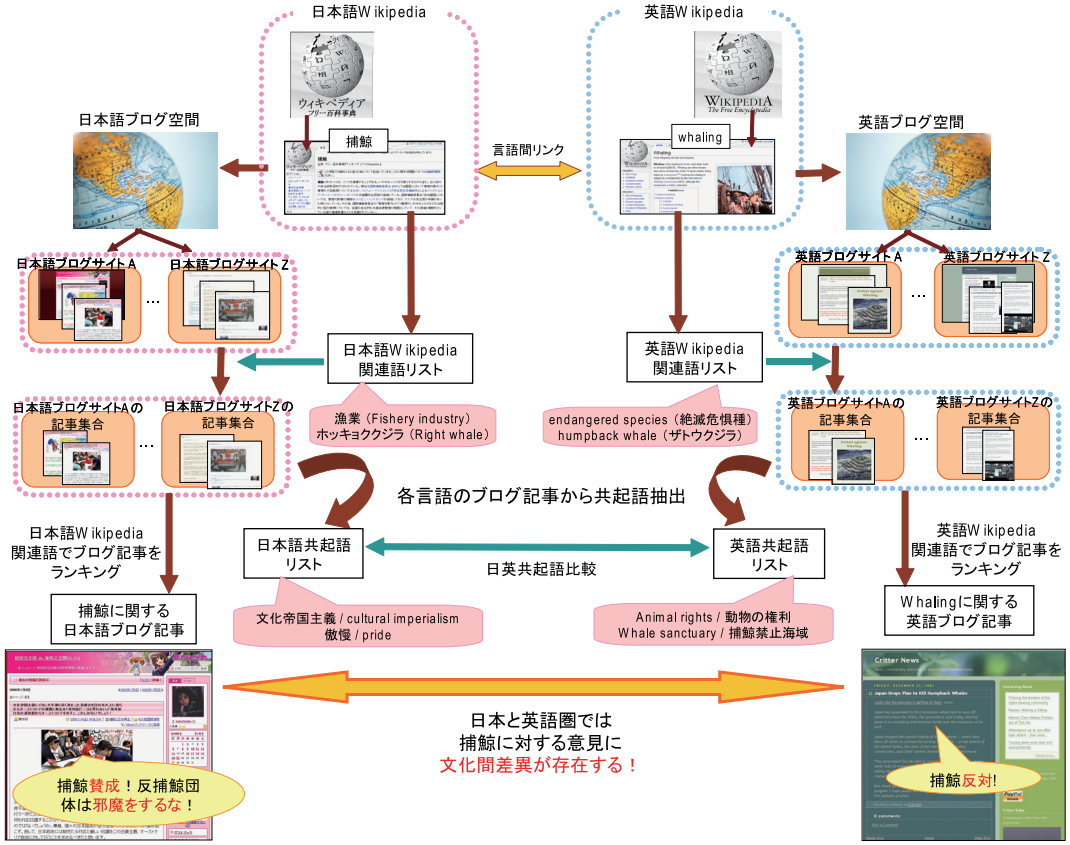


図 1: 二言語対照ブログ分析の全体的枠組み

3.3 ブログ記事からの共起語抽出

本研究では，対照分析の方法として，各言語のブログに出現する共起語を用いる．まず，検索した日本語ブログ記事からは名詞句を抽出し，検索した英語ブログ記事からは一単語，二単語連語，三単語連語を抽出し，それぞれの頻度統計と出現確率を求める．日本語名詞句 X_J の日本語ブログにおける出現確率 $P_J(X_J)$ と，英語一単語・二単語連語・三単語連語 Y_E の英語ブログにおける出現確率 $P_E(Y_E)$ を以下のようにそれぞれ定義する．

$$P_J(X_J) = \frac{X_J\text{の出現頻度}}{\text{対象日本語ブログサイト集合内の総形態素数}}$$

$$P_E(Y_E) = \frac{Y_E\text{の出現頻度}}{\text{対象英語ブログサイト集合内の総単語数}}$$

また，抽出した語句の訳語が相手言語ブログに出現するか調べるために，Wikipedia の言語間リンクを使用して語句の訳語を求める．Wikipedia で語句の対訳を取得できない場合は，英辞郎 *3で語句の対訳を取得する．さらに，抽出した語句の出現率と対訳語句の出現率から，相手言語ブログと比較した出現確率比を求める．本研究では，抽出した日本語名詞句 X_J と X_J の英訳 X_E の出現確率比 $R_J(X_J, X_E)$ と，英語単語・二単語連語・三単語連語 Y_E と Y_E の和訳 Y_J の出現確率比 $R_E(Y_E, Y_J)$ を以下のように定義した．

$$R_J(X_J, X_E) = \frac{P_J(X_J)}{P_E(X_E)},\quad R_E(Y_E, Y_J) = \frac{P_E(Y_E)}{P_J(Y_J)}$$

そして，定義した出現確率比で各言語の共起語をランキングし，それぞれの言語で高い出現確率比の共起語を比較することで，共起語単位でブログ空間におけるトピックの文化間差異発見を支援することができる．

*3 http://www.eijiro.jp/

3.4 ブログサイト・ブログ記事の順位付け

各言語のブログサイト群およびブログ記事群の順位付けにおいては，3.2 節で抽出した Wikipedia 関連語を用いる．ブログ記事は，以下のスコアの降順に順位付けする．

$$PostScore(p) = \sum_t (weight(type(t)) \times freq(t))$$

$weight(type(t))$ は，Wikipedia 関連語 t の種類 $type(t)$ に付与する重みで，$freq(t)$ は，ブログ記事 p 内における Wikipedia 関連語 t の出現頻度である．また，Wikipedia 関連語 t の種類 $type(t)$ がリダイレクトの場合は重みを 3，太字の場合は重みを 2，他エントリリンクの場合は重みを 0.5 とする．また，ブログサイトは，各ブログサイトに含まれるブログ記事のスコアの総和の降順に順位付けする．

3.5 文化間差異発見支援システム

本研究では，同一トピックの日英ブログにおける文化間差異をより発見しやすくするために，文化間差異発見支援システムを作成した．システムの使用方法を図 2 に示す．

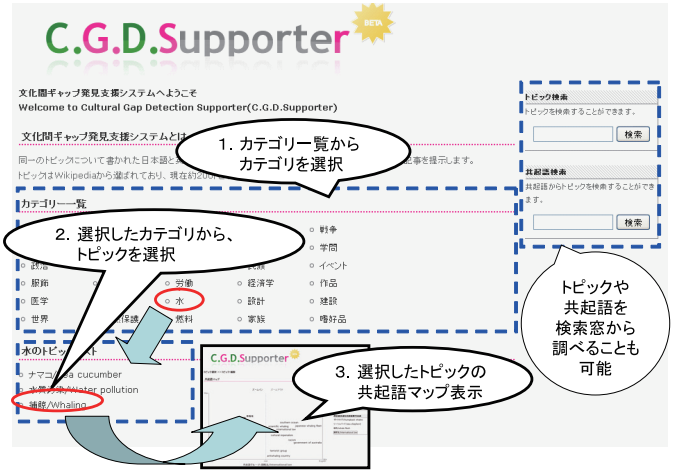
まず，トピックがカテゴリ別に分類されているので，調べたいカテゴリを選択する．カテゴリ情報は，階層構造である Wikipedia の上位カテゴリ約 300 個を用いる．次に，選択したカテゴリのトピックリストが表示されるので，調べたいトピックを選択することで，そのトピックの日英ブログから抽出した共起語を提示する共起語マップを表示することができる．また，検索窓からトピックを検索することで，共起語マップを表示することも可能である．

表 1: 各トピックに関連する日英ブログにおける意見の要約

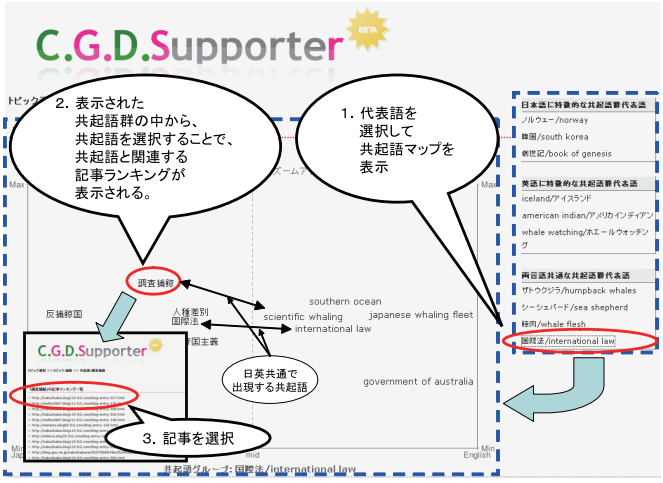
トピック — 概要 日英ブログ間の意見の差異	
(日本語ブログ)	(英語ブログ)
捕鯨（Whaling） — 捕鯨問題において，捕鯨賛成派と捕鯨反対派が対立している．	
多くのブログが捕鯨賛成派．反捕鯨団体を激しく非難している．また，捕鯨について書いているブロガーには，右寄りの考えを持つ人が多くみられた．	多くのブログが捕鯨反対派．特に日本の捕鯨を激しく非難している．また，いくつかのブロガーはホエールウォッチングについて書いている．
臓器移植（Organ transplant） — 治療のために，提供されたドナーの臓器を患者に移植する医療法	
多くのブログは日本の臓器移植法改正の必要性を訴えている．また，いくつかのブログでは，日本の医者によって行われた病気腎移植問題のことに注目している．	多くのブログで，臓器不足という現状から，臓器移植のドナー登録を強く推奨している．また，いくつかのブログでは中国の違法臓器摘出を非難している．
喫煙（Tobacco smoking） — 喫煙することで，人の健康を損なうということで知られている．	
多くのブログで，健康や喫煙マナーの悪さを理由に喫煙に反対しているが，一部のブログは喫煙賛成派である喫煙者のブロガーであった．	多くのブログで，肺がんの原因である喫煙に反対している．
サブプライムローン（Subprime lending） — 近年発生した世界金融危機の大きな原因の一つ	
多くのブログで，米国のサブプライム問題による影響で日本経済が悪化したと指摘．	多くのブロガーが経済学者で，サブプライムローンによって発生した住宅バブルや，現在の金融危機や経営危機の発生原因など考察している．

表 2: Wikipedia から抽出した関連語数，ブログサイト・記事数，ブログ記事中の形態素・単語数

トピック	Wikipedia 関連語数	ブログサイト数	ブログ記事数	総形態素数/総単語数
捕鯨	162 / 174	121 / 239	2232 / 6532	5024966 / 2611942
臓器移植	100 / 231	89 / 206	696 / 1301	995927 / 781476
喫煙	399 / 276	86 / 252	1481 / 400	1323767 / 492727
サブプライムローン	39 / 68	134 / 205	1088 / 1216	980552 / 883450



(a) トピック選択〜共起語マップ表示



(b) 共起語マップ〜関連ブログ記事 URL リスト表示

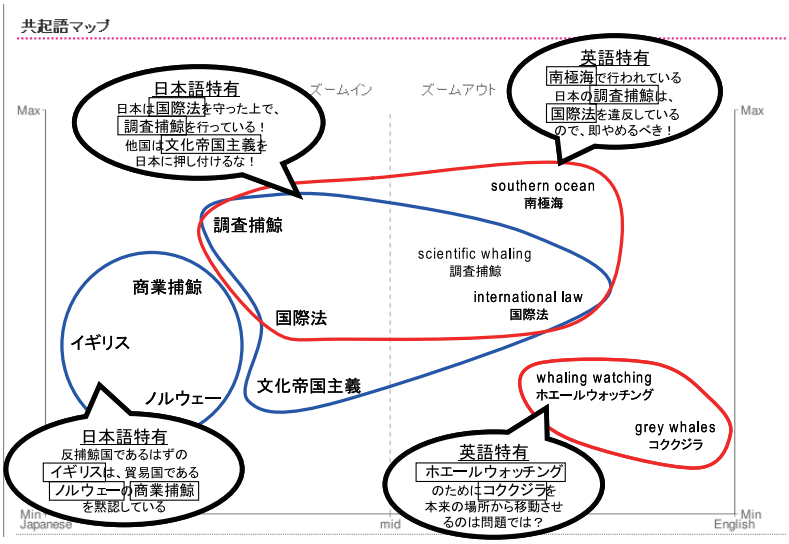
図 2: 文化間差異発見支援システムの使用方法

共起語マップでは，日本語に特徴的な共起語群の代表語，英語に特徴的な共起語群の代表語，両言語共通の共起語群の代表語が提示される．提示された代表語を選択することで，その代表語と関連のある共起語群がマップに表示される．さらに，マップに表示された共起語を選択することで，選択した共起語と関連するブログ記事ランキングが提示される．これによって提示されたブログ記事を分析することで，日英ブログの文化間差異発見の足掛かりとなる．

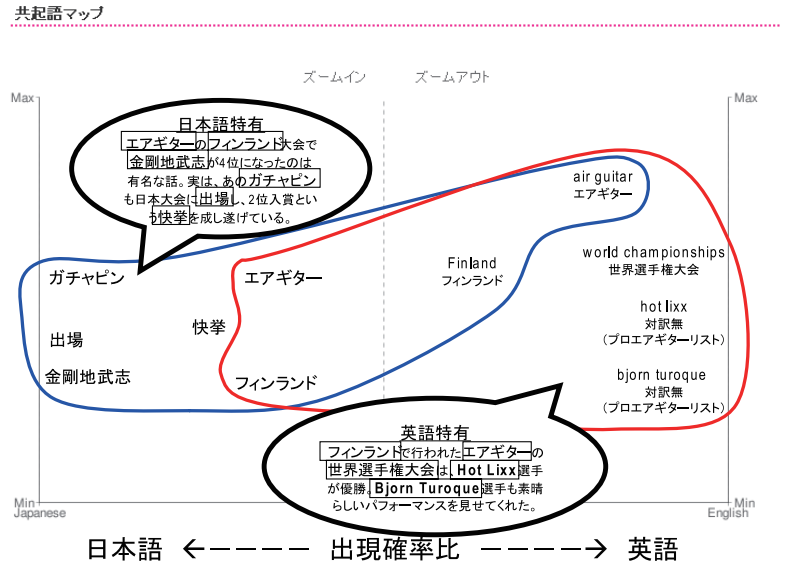
文化間差異発見支援システムを用いて表示した，トピック「捕鯨」と「エア・ギター」の共起語マップ例を図 3 に示す．共起語マップの横軸は，各共起語の出現確率比を表し，縦軸は各共起語の単言語における出現確率を表す．日本語ブログから抽出した日本語共起語 X_J は，座標 $(-R_J(X_J, X_E), P_J(X_J))$ に表示される．このとき，日本語共起語 X_J の英訳 X_E が英語ブログに出現しない場合は，日本語ブログで特徴的な共起語として最も左に表示される．また，英語ブログから抽出した英語共起語 X_E は，座標 $(R_E(X_E, X_J), P_E(X_E))$ に表示される．そして，英語共起語 X_E の和訳 X_J が日本語ブログに出現しない場合は，英語ブログで特徴的な共起語として最も右に表示される．

また，いくつかの共起語は相互的に関係が強く，それらは共通の話題から抽出された共起語群ということがわかる．さらに，片言語のみで特徴的である話題から抽出された共起語群は，縦軸から大きく離れている座標に表示される傾向にある．逆に，両言語で共通している話題から抽出された共起語群は，日英共起語群がそれぞれ縦軸から近い座標に表示されることが多い．

トピック「捕鯨」では，英語ブログで特徴的な共起語は，捕鯨反対の意見をあらわすものが多い．逆に，日本語ブログで特徴的な共起語は，反捕鯨を表明している国を非難している意見をあらわすものが多かった．また，トピック「エア・ギター」では，英語ブログ特有で出現した共起語には，エアギター世界選手権大会で結果を残したプロエアギターリストを賞賛しているものが多かった．逆に，日本語ブログで特徴的な共起語は，エアギターの世界大会で活躍している日本人の話題や，日本の有名なあるキャラクターがエアギターの分野でも活躍していることに驚いているものが多く見られた．このことから，日英ブログから抽出した共起語が日英ブログの文化間差異の発見支援となることがわかった．



(a) 捕鯨 / Whaling



(b) エア・ギター / Air guitar

図 3: 日英ブログから抽出した共起語例を用いた共起語マップ（「捕鯨」,「エア・ギター」）

4. おわりに

本稿では，Wikipedia エントリを用いてトピックに関連する日英ブログサイトを検索し，その記述内容を二言語間で対照分析する方式を提案した．今後は，日英ブログから主観情報・経験情報を多く含む箇所を抽出 [Wiebe 05, 乾 08] することにより，文化間差異測定尺度の高度化に取り組む．また，多言語ブログバースト分析 [福原 07]，複数情報源からのニュースの差異分析 [吉岡 09]，Wikipedia 百科事典，ニュース，ブログといった異種情報源の相補的利用 [佐藤 09] との連携を行う．

参考文献

[福原 07] 福原 知宏, 宇津呂 武仁, 中川 裕志：複数言語間の語彙出現傾向比較による言語横断型ウェブログ関心解析システムの開発, 言語処理学会第 13 回年次大会「大規模 Web 研究基盤上での自然言語処理・情報検索研究」 ワークショップ論文集, pp. 40–43 (2007)

[乾 08] 乾 健太郎, 原一夫：経験マイニング：Web テキストからの個人の経験の抽出と分類, 言語処理学会第 14 回年次大会論文集, pp. 1077–1080 言語処理学会 (2008)

[川場 08] 川場 真理子, 中崎 寛之, 宇津呂 武仁, 福原 知宏：多言語 Wikipedia エントリを用いた特定トピックブログサイト検索と日英対照ブログ分析, 第 22 回人工知能学会全国大会論文集 (2008)

[川場 09] 川場 真理子, 中崎 寛之, 宇津呂 武仁, 福原 知宏：Wikipedia 概念体系を用いた日本語ブログ空間のトピック分布推定, 人工知能学会研究会資料, SIG-SWO (2009)

[中崎 09] 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 (2009)

[佐藤 09] 佐藤 由紀, 中崎 寛之, 川場 真理子, 宇津呂 武仁, 福原 知宏：Wikipedia を知識源とするニュース・ブログ間の相補的ナビゲーション, DEIM フォーラム論文集 (2009)

[Wiebe 05] Wiebe, J., Wilson, T., and Cardie, C.: Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation*, Vol. 39, No. 2-3, pp. 165–210 (2005)

[吉岡 09] 吉岡 真治：NSContrast:世界ニュース比較分析システムの実験的評価, 言語処理学会第 15 回年次大会論文集, pp. 494–497 (2009)