同一トピックの日英ブログサイト検索による二言語対照ブログ分析*

中崎 寛之† 川場 真理子‡ 宇津呂 武仁‡ 福原 知宏§
筑波大学 第三学群工学システム学類†，筑波大学大学院 システム情報工学研究科‡，
東京大学 人工物工学研究センター§

1 はじめに

近年，世界中でブログサービスやブログツールが普及し，各地域の人々がそれぞれインターネット上で個人の意見や評判を発信することが可能になった．それに伴い，様々な情報がブログに記載され，商用ブログ検索サービスを利用することでそれらの情報を取得することができるようになった．具体的なサービスの例として，*Technorati*[1]，*BlogPulse*[2]，*kizasi.jp*[3]，*blogWatcher*[4]などが挙げられる．これらの検索サービスは，巨大なブログ空間の索引付けという観点から見ると，キーワードや評判，時系列変化や人手によって作成されたカテゴリ情報などを索引として用いて，利用者の求めるブログ記事やブログサイトを検索する．また，多言語ブログサービスとしては，*Globe of Blogs* [5]が言語横断ブログ記事検索機能を提供している．他にも，アジア言語ブログの検索機能を提供している *Best Blogs in Asia Directory* [6]や，多言語ブログ記事の分析を行っている *Blogwise*[7]がある．

上記の現状を踏まえた上で，本研究では，新たな多言語ブログ検索機能の実現へ向けて，**同一トピックについてまとまった規模の記述が書かれているブログサイトを，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．英語ブログには，日本語ブログに記述されていない文化独特の意見や評判が書かれていることが多い．本稿の目的を達成すれば，そのような意見や評判を知ることができ，特定トピックに対する意見の国間差異を見つけることができる．

上記の目的を達成するために，巨大なブログ空間へのアクセスを実現するにあたって，あらゆる事柄が詳細に体系化された知識体系である Wikipedia とブログサイトを対応づけるアプローチをとる [川場 08]．Wikipedia は誰でも自由に情報を書き込むことが可能な巨大ウェブ百科事典として知られており，様々な分野に関する詳細な情報が記載されている．また，トピックに対するブログサイトの有無などを知ることによって，現存するブログ空間における話題の分布の傾向を把握することが容易に実現できる．さらに，検索対象のブログの単位をブログサイトとすることで，ブログ空間において，個々の記事よりも大きいブログ著者の単位で索引付けすることができる．

本稿の目的は，同一トピックについて書かれている日英ブログサイトをそれぞれ検索し，それらの記述内容を対照分析することにある．この目的を達成するために，上記で述べた検索の枠組みを踏まえた上で，Wikipedia の言語間リンクを使用する．言語間リンクを辿ると，特定トピックの日本語 Wikipedia エントリから，同一トピックの英語 Wikipedia エントリを読むことができる．つまり，英語 Wikipedia のエントリを英語ブログサイトへの索引とすれば，同一トピックの英語 Wikipedia エントリに対応づけた英語ブログサイトを取得することができ，本稿の目的を達成することができる．

2 Wikipedia

Wikipedia とは多くの人が自由に書くことができるインターネット上の巨大な辞書のことであり，日本語で約 45 万，英語で約 220 万のエントリ (2008 年 1 月現在) がある．大きな特徴として，Wiki を利用して作られており，だれでも自由に情報を書き込むことができる．さらに，11 のメインカテゴリ以下にサブカテゴリ，エントリが連なる，巨大な木構造になっている．また，カテゴリが木構造のノードにあたり，エントリが木構造の葉に相当する．図 1 に示すように，日本の電気通信事業者カテゴリというノードの下にさらにサブカテゴリがノードとしてつながっており，さらにそのカテゴリの下の NTT グループサブカテゴリの下には日本電信電話エントリが葉となってつながっている．

また，Wikipedia は多くの言語で書かれており，言語間リンクを辿ることで他の言語で書かれたエントリを読

---

*Cross-Lingual Blog Analysis based on Japanese/English Blog Distillation

†Hiroyuki Nakasaki, College of Engineering Systems, Third Cluster of Colleges, University of Tsukuba
‡Mariko Kawaba, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
§Tomohiro Fukuhara, Research into Artifacts, Center for Engineering, University of Tokyo

[1] http://technorati.com/
[2] http://www.blogpulse.com/
[3] http://kizasi.jp/（日本語のみ）
[4] http://blogwatcher.pi.titech.ac.jp/（日本語のみ）
[5] http://www.globeofblogs.com/
[6] http://www.misohoni.com/bba/
[7] http://www.blogwise.com/

図 1: Wikipedia の構造

むことができる．本稿の実験に用いた日本語キーワードに対応する英語キーワードは Wikipedia の言語間リンクの情報を使用した．

3 同一トピックの日英ブログサイト検索

3.1 検索手法

本研究ではまず，Wikipedia の中のある特定のトピックから，そのトピックについての意見や評判などの情報が書かれている日英ブログサイトを探し，対応づける．しかし，現在のブログ検索サービスでは，被リンク数の多い人気ブログサイトの記事から優先的に検索されるために，被リンク数は多くないが，特定トピックについて濃い情報を載せているブログサイトが検索されにくい．本研究の目的を達成するためには，トピックについて濃い情報を載せているブログサイトの集合を得る必要がある．よって，被リンク数の多い，人気度の高いブログサイトを優先的に検索するのではなく，検索トピックについて多く述べられているブログサイトを優先的に検索する必要がある．そこで，本稿では，検索トピックがブログサイトにどれだけ出現しているかで検索トピックについて述べられているブログサイトかどうかを判定するという手法 [川場 08] を用いる．つまり，**検索トピックの出現数が多いブログサイトを検索する**というアプローチをとる．具体的には図 2 に示すように，

> **通常の検索方法でブログサイトを検索した後，検索されたブログサイト集合を検索トピックの出現数が多い順にソートする．**

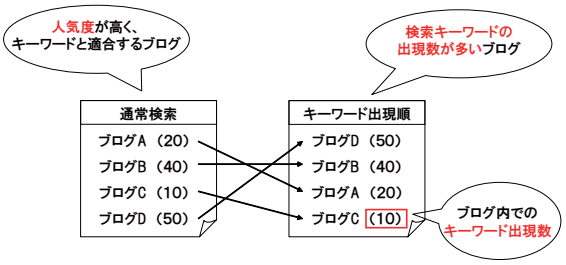

図 2: 特定トピックに一致するブログの検索手法

3.2 実験手順

ブログサイトを検索するために，本実験では日本語ブログの検索には，Yahoo!Japan 検索 API [8]を，英語ブログの検索には米 Yahoo!検索 API [9]を利用し，日本語ブログでは大手 11 社 [10]，英語ブログでは大手 12 社 [11] 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，複数のドメインを一度に指定して検索し，1,000 件の記事を取得する [12]．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，1 キーワードあたり約 200 前後のブログサイトを取得することができた．また，提案手法ではこれらのブログサイトをキーワードの出現数順に並べ替えるが，並び替える前の，API の出力順にブログサイトをランキングしたものをベースラインとした．ここで，日米の Yahoo!検索 API で，ブログサイトをドメイン指定してキーワードを検索した際に求められる検索結果の数を検索キーワードの出現回数とした．

検索キーワード

本研究では，あるトピックに対する日英のブログサイトの記述内容を，二言語間で対照分析するというタスクに対して，本研究で提案するブログサイト検索手法を適用する．そこで，評価実験に使用した検索キーワードとしては，Wikipedia のエントリのタイトルを対象として，日本に関する幅広い分野のトピックで，かつ，日本語・英語共にある程度の数のブログ集合が得られるようなトピックを約 60 選定した．表 1 に選定したキーワードを

[8] http://developer.yahoo.co.jp/search/
[9] http://developer.yahoo.com/search/
[10] 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
[11] 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
[12] 本稿の評価実験では，ベースラインとの比較において，ベースラインにおける順位付けを利用する必要があるという制約から，各ドメインごとに検索を行うのではなく，複数ドメインを一度に指定して検索を行っている．現在，この検索とは別に，各ドメインごとに 1,000 記事ずつ検索して，これらの和集合に対して，提案手法の順序並べ替えを行うという方法で評価実験を行っている．

表 1: 検索に使用したキーワード

分野	検索キーワード
アニメ	ドラえもん，ポケモン，鉄腕アトム，機動戦士ガンダム，ドラゴンボール，新世紀エヴァンゲリオン，セーラームーン
音楽	パフィー，X Japan，Dir en grey，ジャズ，交響曲
動物	犬，猫，ハムスター，ジャイアントパンダ，チワワ
企業	ソニー，カシオ，任天堂，ホンダ，トヨタ，三洋電機，キャノン
商品	PS3，PSP，iPod，Wii，ニンテンドー DS
歴史・文化	自衛隊，原爆，寿司，自民党，天皇，富士山
社会問題	靖国神社，年金，NOVA，捕鯨，テロ
施設	博物館，水族館，ミュージカル，遊園地，ディズニーランド
スポーツ	ボクシング，亀田興毅，亀田大毅，プロ野球，中村紀洋，イチロー，福留孝介，松坂大輔，井川慶，相撲，プロレス，K-1

示す．

これらの約 60 キーワードの内，「ドラゴンボール」，「Wii」，「新世紀エヴァンゲリオン」，「靖国神社」，「捕鯨」[13]の 5 キーワードを選び，それぞれ，上位 30 位と以下等間隔に 30 ブログサイトをサンプリングし，手動で評価した．また，手動評価の際，特定トピックについてある一定数以上のブログ記事があれば正解とし，一定期間特定トピックについて書かれているということは考慮していない．

4 日英ブログの二言語対照分析

本稿では，収集した日英のブログサイトから，それぞれ前節の実験と同じ方法で 60 個サンプリングし，内容を分析し比較した．以下にその結果を述べる．

分析の結果，商品のような物に関するキーワードと，社会問題のような意見に関するキーワードではブログの内容の違いにも差が出るということがわかった．具体的にはドラゴンボール，Wii，新世紀エヴァンゲリオンのような商品となるキーワードで検索されたブログは日本と欧米の社会的な興味や関連商品の需要の違いがあることから，記事に書かれている内容に違いが見られた．ドラゴンボールを例に挙げると，日本ではアニメの放送も終了していることやウェブ上の動画に対する規制が厳しいこともあり，日本語ブログにおいてはドラゴンボールに関連したゲームやカードダスの商品について記載している記事が多くみられた．しかし，英語ブログでは，ウェブ上の映像に対する規制が日本ほど厳しくないこともあるためか，ドラゴンボールのアニメ動画を記事に載せているブログがいくつかみられた．一方，社会問題のようなキーワードで検索されたブログでは，日本語ブログと英語ブログで全く逆の意見が見られた．日本語のブログでは国粋主義，右寄りのブロガーによって書かれた日本の行為に対する肯定的意見の多いブログが多く，英語のブログでは反日のブロガーによって書かれた，日本の行為に対する否定的な意見の多いブログが多く見られた．この傾向は，捕鯨と靖国神社の両キーワードでそれぞれ強く見られた．キーワードの詳細とブログの内容の詳細を表 2 に示す．

[13]捕鯨に関しては，提案手法によって収集できたブログ数が 30 ブログ前後であり，また捕鯨について述べられていたブログ数が 30 ブログ中 3 ブログしか現れなかったために，評価結果のグラフは省略した．しかし，各ドメインあたり 100 ブログ記事を検索した場合，約 400 のブログサイトを取得することができ，また，上位に多くの捕鯨に関するブログを確認することができた．

5 関連研究

本研究の関連研究には，ある話題に関する意見を賛成と反対に分けてウェブから収集する研究 [井上 07] がある．この研究は，ある話題に関する賛否両論の集約と可視化を目的としている．本研究では，収集した日英ブログにおける賛否両論の集約を自動化することで，この研究を適用することができる．また，多言語ブログに関する研究としては，日韓中英のブログ内で，キーワードのバーストの時系列の変化を各言語間で調べるという研究 [福原 07] がなされている．本研究では，ブログの内容を見ているため，キーワードのバーストの時系列の変化を調べることは行っていない．

また，Wikipedia に関する研究として図書館の分類体系と Wikipedia カテゴリの対応付けを行う研究 [田村 07] がある．その他に，Wikipedia から固有表現を取出す研究 [Cucerzan07, Kazama07]，Wikipedia の言語間リンクを利用して多言語対訳辞書を作成するという研究などがなされている [新井 08]．本稿の日英ブログ対照分析に用いた対訳は，この手法 [新井 08] と同様，Wikipedia の言語間リンクを使用している．

6 まとめと今後の課題

本稿では Wikipedia とブログ集合の対応付けのために，トピックの出現回数の多いブログを検索することでブログ集合を検索する検索実験行った結果を報告した．実験の結果から現在の検索手法の改善すべき点を述べた．また，検索によって集めたブログ集合を用いて日英ブログの対照分析を行い，日英ブログを同一キーワードで検索した際に，検索キーワードの持つ属性によって日本語と英語のブログの内容に違いがみられることがわかった．

しかし，本稿で行った検索ではまだ，ノイズも多く混入してしまい，Wikipedia のトピックに対応するブログを十分に収集できているとは言いがたい．また，Wikipedia エントリのタイトルのみを使用して検索することだけでは不十分であると考えられる．より精度良く検索を行うために，今後 TREC の Blog Distillation タスク [Mac-

表 2: 日英ブログの言語対照分析

日本語トピック名 (英語トピック名)	簡単な説明 現象や意見の違い (日本語ブログ)	(英語ブログ)
ドラゴンボール (Dragon Ball)	日本の漫画作品．40ヶ国以上で翻訳されている．	
	多くのブログがドラゴンボールカードダスについて書かれていた．また，ゲームのレビューなどがいくつかあったが，著作権違反になるために，動画はほとんど見られなかった．また，ドラゴンボールのキャラクターの形を模した弁当の画像を多く載せているブログもいくつかあり，同人誌などを作成しているブログも数件みられた．	ドラゴンボールカードダスについてはほとんど述べられていなかった．また，ゲームのレビューはあまり多くないが，ゲームやアニメの動画へのリンクなどが張ってあるブログが多く見られた．また，ドラゴンボールのファンであり，ドラゴンボールに関すること全般に渡って書かれているブログや，訪問者へドラゴンボールに関するアンケートを行っているブログもいくつかみられた．
Wii (Wii)	任天堂から発売された据え置き型ゲーム機．世界中で販売されている．	
	Wii の改造などは日本の法律違反になるために，Wii の改造について述べられているブログは全く見られなかった．ゲームタイトルの一覧があり，それぞれに詳細な感想，評価などが述べられているブログが多くみられた．また，2007 年 10 月に発売された Wii Fit を使用したダイエットの記録をつけたブログなども多く見られた．関連商品へのアフィリエイトも多く見られた．	いくつかのブログに Wii で自作のソフトウェアを動かす方法や，Wii の改造について述べていた．また，多くのブログにゲームの動画や Wii の写真などが載せられていて，ゲームについての感想などが書かれているブログも数個見られた．
新世紀エヴァンゲリオン (Neon Genesis Evangelion)	日本のアニメ．英語版がヨーロッパやラテンアメリカ，アジアなどで放送されている．	
	多くのブログでアニメや映画の感想が書かれていた．また，パチンコを趣味とする人のブログでエヴァンゲリオンのパチンコについての感想や攻略法を書いたブログが見られた．また，エヴァンゲリオンの関連商品にアフィリエイトを張った，スプログもいくつか見られた．著作権の問題からか，動画が載せられているブログはあまり見られなかった．	多くのブログにアニメの動画やアニメの画像を編集した動画が載せられていた．また，自作のデスクトップ画像の配布を行っているブログややフィギュアの画像を多く載せているブログなどが見られた．また，いくつかのブログにアニメの感想などが書かれていた．エヴァンゲリオンのパチンコをしたブログは全く見られず，また，スプログもほとんど現れなかった．
捕鯨 (Whaling)	捕鯨に，反対か，賛成かの意見が書かれたブログが見られた．	
	多くのブログが日本の行為に対し肯定的であり，捕鯨に賛成していて，反捕鯨団体を激しく非難しているブログなどもいくつか見られた．また，日本の新聞，雑誌，テレビなどのメディアが捕鯨に関するニュースをどのように伝えているかという分析を客観的に行っているブログも見られた．捕鯨に関する記事を書いているブロガーは国粋主義，右寄りのブロガーが多くみられた．	多くのブログが日本の捕鯨に反対していて，反捕鯨運動を呼び掛けているブログなども多く見られた．また，ホエールウォッチングをしているブログがいくつか見られた．
靖国神社 (Yasukuni Shrine)	東京にある神道の神社．国会議員や内閣総理大臣が参拝することが問題になっている．	
	全体的に靖国神社参拝に肯定的なブログが多くみられた．また，会合など開き，靖国参拝を呼び掛けているブログも多く見られた．また，靖国神社についてのブログでは国粋主義，右寄りの考えを持つブロガーが多く見られ，意見性の強い記事も多く見られた．	大半のブログが日本の国会議員の靖国神社参拝に否定的な意見を述べていた．また，日本人の過去の歴史に対する意識を批判しているブログが見られた．靖国神社参拝だけではなく，靖国神社そのものに関する批判も多く見られた．中には靖国神社の事を「Yasukuni War Shrine」と表記しているブログも見られた．

donald07] の成果なども取り入れて，ノイズ除去，検索質問拡張などを行っていく必要がある．

また，日英ブログの対照比較分析においても，本稿の実験で行った商品と社会問題の 2 つの種類のキーワードでの分析だけでは不十分であり，今後様々な種類のキーワードを検索した結果の分析を行っていく予定である．

参考文献

[新井 08] 新井嘉章, 福原知宏, 増田英孝, 中川裕志：Wikipedia を用いた多言語ブログ検索のための訳語抽出, 情報処理学会第 70 回全国大会講演論文集, 情報処理学会 (2008).

[Cucerzan07] Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 708–716 (2007).

[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).

[井上 07] 井上結衣, 藤井敦：時事問題に関する賛否両意見の収集, 情報処理学会研究報告, Vol. 2007, No. (2007–NL–181), pp. 93–98 (2007).

[川場 08] 川場真理子, 宇津呂武仁, 福原知宏：Wikipedia エントリに対応するトピックのブログサイト検索, 言語処理学会第 14 回年次大会論文集 (2008).

[Kazama07] Kazama, J. and Torisawa, K.: Exploiting Wikipedia as External Knowledge for Named Entity Recognition, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 698–707 (2007).

[Macdonald07] Macdonald, C., Ounis, I. and Soboroff, I.: Overview of the TREC-2007 Blog Track, *Proceedings of thej TREC-2007 (Notebook)*, pp. 31–43 (2007).

[田村 07] 田村悟之, 清田陽司, 増田英孝, 中川裕志：図書館における自動レファレンスサービスシステムの実現 Web 上の二次情報と図書館の一次情報の統合, 情報処理学会研究報告, Vol. 2007, No. (2007–FI–179), pp. 1–8 (2007).