特 集

多言語Wikipediaエントリを知識源とする特定トピックの日英ブログサイト検索と日英対照ブログ分析

Japanese/English Blog Distillation and Cross-Lingual Blog Analysis with Multilingual Wikipedia Entries as Fundamental Knowledge Source

中崎 寛之 Hiroyuki Nakasaki
株式会社 NTT データ
NTT DATA CORPORATION

川場 真理子 Mariko Kawaba
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

横本 大輔 Daisuke Yokomoto
筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba
http://nlp.iit.tsukuba.ac.jp/

宇津呂 武仁 Takehito Utsuro
（同　　上）
utsuro'at'iit.tsukuba.ac.jp, http://nlp.iit.tsukuba.ac.jp/

福原 知宏 Tomohiro Fukuhara
独立行政法人 産業技術総合研究所 サービス工学研究センター
Center for Service Research, National Institute of Advanced Industrial Science and Technology

keywords: blog, topic analysis, Wikipedia, blog distillation, cross-lingual blog analysis

Summary

The overall 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. In this paper, we take an approach of collecting blog feeds rather than blog posts, mainly because we regard the former as a larger information unit in the blogosphere and prefer it as the information source for cross-lingual blog analysis. In the blog feed retrieval procedure, we also regard Wikipedia as a large scale ontological knowledge base for conceptually indexing the blogosphere. The underlying motivation of employing Wikipedia is in linking a knowledge base of well known facts and relatively neutral opinions with rather raw, user generated media like blogs, which include less well known facts and much more radical opinions.

In our framework, first, in order to collect candidates of blog feeds for a given query, we use existing Web search engine APIs, which return a ranked list of blog posts, given a topic keyword. Next, we re-rank the list of blog feeds according to the number of hits of the topic keyword as well as closely related terms extracted from the Wikipedia entry in each blog feed. We compare the proposed blog feed retrieval method to existing Web search engine APIs and achieve significant improvement. We then apply the proposed blog distillation framework to the task of cross-lingually analyzing multilingual blogs collected with a topic keyword. Here, we cross-lingually and cross-culturally compare less well known facts and opinions that are closely related to a given topic. Results of cross-lingual blog analysis support the effectiveness of the proposed framework.

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]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

*1 http://technorati.com/
*2 http://www.blogpulse.com/
*3 http://kizasi.jp/(日本語のみ)

た，多言語ブログサービスとしては，*Globe of Blogs*[*4]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* [*5]や，多言語ブログ記事の分析を行っている *Blogwise*[*6]がある．

ここで，本論文では，日英ブログの分析を目的としてブログの検索を行う．特に，本論文では，個々のブログ記事ではなく，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に注目する．そして，そのような専門的内容を含むブログサイトを選択的に検索する手法を提案する．本論文において実現をめざす手法と比較すると，既存の検索エンジン API を用いたブログ検索においては，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傾向にある．したがって，既存の検索エンジン API を用いた場合は，被リンク数は多くないが，特定のトピックについて詳細な情報を載せているブログサイトが検索されにくい．この問題に対して，本論文の手法では，特定のトピックについての詳細な情報を含むブログサイトを選択的に検索することを実現するために，各トピックについての Wikipedia エントリ中の記述を知識源として利用する (3 章)．特に，本論文では，日英ブログの分析の準備段階として，日英両言語のブログサイトの検索を行うが，その際には，Wikipedia における日英両言語のエントリを知識源とする．そして，トピック名がタイトルである Wikipedia エントリを知識源として，トピックに密接に関連する用語を抽出し，それらの関連語がより多く含まれるブログサイトを検索するという手法を用いる．実際に，本論文の手法を既存の検索エンジン API と比較し，ブログサイトの検索性能において本論文の手法が優れていることを示す (4 章)．

さらに，本論文では，日英両言語において，詳細な内容を記述したブログサイトが一定数存在するトピックを対象として，上述の手法により検索したブログサイト，および，各ブログサイトのブログ記事の内容を，日英各言語において分析するとともに，二言語間でブログ中の内容の比較を行う．本論文では特に，社会現象および社会問題に関するトピックに焦点を当てて，日英間でのブログ分析を行った結果を事例として示す．そして，実際に，各トピックごとに興味深い言語間差異を観測した結果を示す (5 章)．

本研究の全体的枠組みを図 1 に示す．まずトピック名である日英 Wikipedia エントリのタイトルを検索語として，日英ブログサイトを収集する．次に，トピック名についての日英 Wikipedia エントリから関連語 (以下，本論文では，Wikipedia 関連語と呼ぶ) を抽出し，抽出した Wikipedia 関連語を用いて，検索したブログサイト集合からトピックに関わりのあるブログ記事を選別する．そして，Wikipedia 関連語を用いてブログサイト集合およびブログ記事集合を自動順位付けした後，上位に順位付けされたブログサイトおよびブログ記事を人手で分析する．この枠組みにおいて，人手によって，ブログ中の記述を日英二言語間で比較・対照分析することにより，日本および英語圏の間で，トピックに関する関心・意見の共通点・差異がどの程度存在するのかを発見することが容易になる．

本研究は，近未来チャレンジ「Wikipedia マイニング」[中山 09] に参画している．このチャレンジでは，ここ数年爆発的に急成長してきたオンライン百科事典である Wikipedia を解析対象として，情報検索，自然言語処理，人工知能等の幅広い分野の応用において有用な知識を抽出することを目的としている．さらに，概念同士の関連性を連想関係として抽出する技術，概念間の意味関係を抽出する技術，および，多言語百科事典である Wikipedia からの対訳関係の抽出，の 3 点に焦点を当てるとともに，情報検索・文書分類等の応用においてこれらの知識の有用性を示すことを目指している．これらの知識の中でも，特に，概念間の連想関係については，情報検索の高度化における有用性を示すとしている．これに対して，本研究においては，情報検索タスクの一つとして，ブログサイトおよびブログ記事の検索の際の順位付けにおいて，Wikipedia エントリから抽出した関連語を用いた順位付けの性能が既存の検索エンジン API の順位付けの性能を上回ることを示す．ここで，本研究で用いる Wikipedia 関連語は，Wikipedia エントリにおける概念間の連想関係ととらえることができるので，本研究は，まさに，Wikipedia マイニングによって抽出した知識が情報検索タスクにおいて有用であることの実証例と位置付けることができる．

また，[中山 09] においては，Wikipedia マイニングによって抽出した対訳辞書を利用して異言語横断文書検索を実現することが課題として提示されている．これに対して，本研究においては，Wikipedia エントリのタイトル間の言語間リンクのみを翻訳知識として用い，日英二言語において Wikipedia 関連語を用いたブログサイト・ブログ記事の検索を行うことによって，日英二言語のブログの間で興味深い言語間差異が観測できることを示す．このように，本研究は，ウェブ上における重要な多言語情報源の一つである多言語ブログを検索対象として，Wikipedia に含まれる多言語知識を特に加工することなくそのまま用いることにより，異言語横断文書検索が容易に実現でき，さらに，検索結果を対象として言語間差異の対照分析が実現できることの実証例としても位置付けることができる．

2. 分析対象トピック

まず，5 章で述べる日英対照ブログ分析の対象とするトピックとしては，日英両言語のブログ空間において，

*4 http://www.globeofblogs.com/
*5 http://www.misohoni.com/bba/
*6 http://www.blogwise.com/

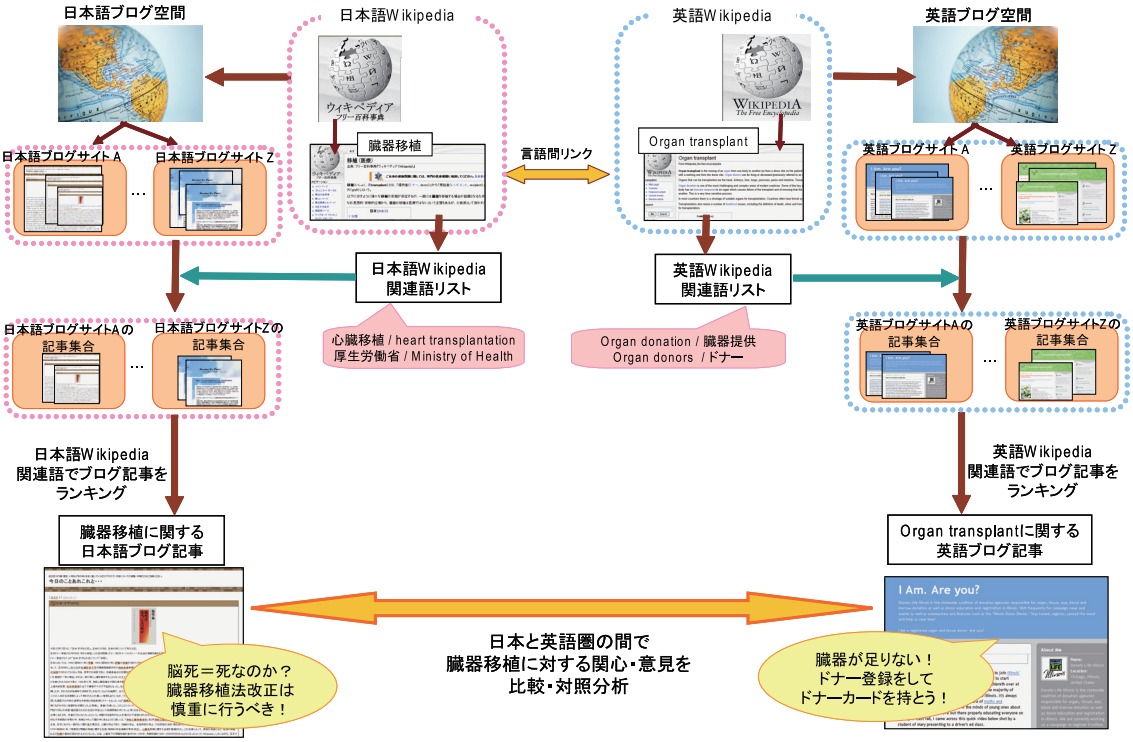


図 1 日英対照ブログ分析の全体的枠組み

表 1 各トピックごとの Wikipedia 関連語数およびブログサイト・記事数 (日本語/英語)

評価用トピック	Wikipedia 関連語数	ブログサイト数	ブログ記事数
アルコール依存症	60 / 161	183 / 100	2158 / 5185
リストラ	70 / 20	102 / 102	2640 / 6944
著作権侵害	88 / 108	56 / 99	1195 / 4448
捕鯨	164 / 169	435 / 99	8026 / 6657
臓器移植	94 / 264	90 / 97	1691 / 1955
喫煙	399 / 304	358 / 86	11992 / 952
サブプライムローン	39 / 68	382 / 95	6470 / 1400

そのトピックについて詳細な記述を掲載しているブログサイトが十分な数存在する可能性が高いトピックが望ましい．そのための手がかりとして，予備調査として，日本語・英語それぞれのブログ空間におけるトピック名のヒット数の範囲と，詳細な記述を掲載しているブログサイトの有無との相関を分析した結果，ブログ空間におけるヒット数が 10,000 以上であれば，詳細な記述を掲載しているブログサイトが存在する可能性が比較的高いことが分かった．実際に，日英両言語の間で言語間リンクを介して対訳関係にあるエントリが存在するエントリのうち，日本語 Wikipedia エントリのタイトルが日本語ブログ空間中で 10,000 以上のヒット数を持ち，かつ，英語 Wikipedia エントリのタイトルも英語ブログ空間中で 10,000 以上のヒット数を持つエントリは，約 6,000 個存在した．本論文では，そのうち，特に，社会現象および社会問題に関するトピックに焦点を当てて，約 50 個のトピックを選定し，日英ブログサイト・ブログ記事の検索・自動選定を行った．また，各トピックについて詳細な内容を記述しているブログサイトの分析を行うとともに，日英二言語間でブログ中の記述内容の対照分析を行った．それらのトピックのうち，本論文では，「アルコール依存症 (alcoholism)」，「リストラ (restructuring)」，「著作権侵害 (copyright infringement)」，「捕鯨 (whaling)」，「臓器移植 (organ transplant)」，「喫煙 (tobacco smoking)」，「サブプライムローン (subprime lending)」の計 7 トピックについての分析結果について述べる．

また，日英両言語を対象として，4 章で述べるブログサイト検索手法を評価するためのトピックとしては，上記 7 トピックのうち，3 トピックを選定した*7．

*7 これまでに，提案手法によって日本語および英語ブログサイト・ブログ記事を自動順位付けした結果を用いて日英各言語のブログ分析を行っているが，4 章における評価に用いた 3 トピック以外のいずれのトピックにおいても，定性的評価の範囲では，検索エンジン API による順位付けと比較して，提案手法によって一定の改善を達成できている．

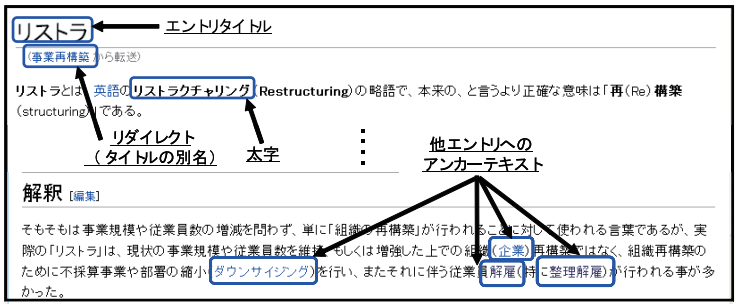


図 2 Wikipedia エントリおよび関連語の例

3. Wikipedia を用いた関連語の収集

3·1 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 64 万，英語で約 314 万のエントリ (2010 年 1 月現在) がある．さらに，10 個程度の主要カテゴリ以下にサブカテゴリ，エントリが連なる，巨大なグラフ構造になっている．また，カテゴリがグラフ構造の節にあたり，エントリが節内に列挙されている．

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むことができる．これまでに，すでに，世界の主要な言語版の Wikipedia が存在するため，十分な種類・数のブログが書かれている言語を対象として本論文の手法を適用することは比較的容易である．また，他の知識源と比較した場合の Wikipedia の最大の利点として，日常的に，新たなエントリの作成と記述の更新が行われており，ブログにおける分析対象となり得る主要なトピックが網羅されている点が挙げられる．

3·2 Wikipedia 関連語の収集

トピック名がタイトルである各言語の Wikipedia エントリを知識源として，トピック名に密接に関連する Wikipedia 関連語を収集する．特に，本論文においては，各エントリのリダイレクト，各エントリ本文中の太字，および，本文中における他エントリへのリンクのアンカーテキストを Wikipedia 関連語として収集する．Wikipedia エントリ「リストラ」の場合について，エントリのスナップショットの抜粋，および，Wikipedia 関連語の例を図 2 に示す．この例の場合は，エントリタイトル「リストラ」の別名であるリダイレクト「事業再構築」で検索を行った結果，エントリタイトル「リストラ」の下に，リダイレクト「事業再構築」が表示され，エントリとしては「リストラ」の本文が提示されている．その他，エントリ中の太字「リストラクチャリング」，および，他エントリへのリンクのアンカーテキスト「企業」,「ダウンサイジング」,「解雇」,「整理解雇」が提示されている．本論文において評価対象とする各トピックについて，収集した Wikipedia 関連語数を表 1 に，Wikipedia 関連語の抜粋を表 2 に，それぞれ示す．

4. Wikipedia エントリに対応するブログサイトの検索

4·1 ブログサイトの収集

まず，分析対象となるブログサイトの候補を収集するために，本論文では，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]をそれぞれ利用する．ただし，日本語ブログホスト大手 8 社[*10]，および，英語ブログホスト大手 4 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，1000 件の記事を取得する．しかし検索エンジン API を用いた検索ではブログ記事単位の検索になるので，ブログ記事検索後，ブログサイト単位にまとめた．その結果，1 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．

4·2 ブログ記事の選別

次に，収集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選定する．手法としては，3·2 節において収集した Wikipedia 関連語のいずれかが出現する各言語のブログ記事を選定する．本論文で評価対象とした各トピックについて，以上の手法により収集したブログサイト数，および，収集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を表 1 に示す．

4·3 ブログサイト・ブログ記事の順位付け

次に，トピックについて詳細な記述を多く含むブログサイトおよびブログ記事をより上位に順位付けするために，3·2 節で抽出した Wikipedia 関連語を用いて，ブログ記事およびブログサイトにスコアを付与する．まず，ブ

*8 http://www.yahoo.co.jp/
*9 http://www.yahoo.com/
*10 FC2.com, yahoo.co.jp, rakuten.ne.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, hatena.ne.jp
*11 blogspot.com, livejournal.com, typepad.com, wordpress.com

表 2 各トピックの Wikipedia 関連語の抜粋 (リダイレクト / 太字 / 他エントリへのリンクのアンカーテキスト)

日本語トピック名 (英語トピック名)	日本語	英語
アルコール依存症 (alcoholism)	アルコール依存，慢性アルコール中毒，酒乱 / アルコール中毒 / 日本酒，飲酒運転，アルコール飲料，ビール，ウェルニッケ脳症，肝炎	Alcohol addiction, Alcoholic abuse, Alcohol misuse, Drinking problem, Alcholic / abuse, recovery, problem use, heavy use / Blood alcohol content, Domestic violence,
リストラ (restructuring)	事業再構築 / リストラクチャリング / アウトソーシング，バブル崩壊，レイオフ，ワークシェア，解雇，終身雇用，退職勧奨	Corporate restructuring / / Bankruptcy, Compromise agreements, Layoff, Outsourcing, Spin-out, Voluntary Redundancy
著作権侵害 (copyright infringement)	/ 依拠性，創作的，類似性 / フェアユース，レコード輸入権，意匠権，海賊版，告訴，裁判所，実用新案権，著作権，著作物，特許権，日本音楽著作権協会	Copyright violation, Illegal copying, Pirated music, Unauthorised copying, Unlawful copying / Bootleg recording, piracy / Copyright misuse, EU Copyright directive, Intellectual Property, Peer-to-peer, Plagiarism
捕鯨 (whaling)	捕鯨国 / 「鯨塚」,「鯨墓」 / ゴンドウクジラ，ザトウクジラ，鯨漁取締規則，鯨料理，国際捕鯨委員会，捕鯨問題	Whale fishing, Whalehunter / Subpopulations, Subspecies / Animal intelligence, Animal welfare, Endangered species, Greenpeace, Humpback Whale, Meat and bone meal
臓器移植 (organ transplant)	移植手術，生体肝移植 / 心臓死移植，臓器，脳死移植 / 臓器提供意思表示カード，日本臓器移植ネットワーク，アイバンク，移植コーディネーター，骨髄移植，再生医学	Heart Transplant, Kidney Transplantation, Organ & Tissue Donor, Transplanted organs / Websites about Illegal Organ Procurement, living donors / Brain death, Eye bank, Human Tissue Authority,
喫煙 (tobacco smoking)	ヘビースモーカー，愛煙家，煙草 / / たばこ特別税，咽頭ガン，気管支喘息，喫煙席，禁煙，携帯灰皿，紙巻きタバコ，受動喫煙，日本たばこ産業	Cigarette smoking, Nicotine addiction / / Cigar, Nicotine gum, Passive smoking, Second-hand smoke, Smoking culture, Smoking ban
サブプライムローン (subprime lending)	サブプライム，サブプライム問題 / / バブル経済，金融危機，住宅ローン，消費者金融，連邦準備制度理事会，不動産担保証券	Non-prime loans, Subprime mortgage / near-prime / Bankruptcies, Mortgage loan, United States Department of Housing and Urban Development

ログ記事 p のスコアとして，以下を用いる．

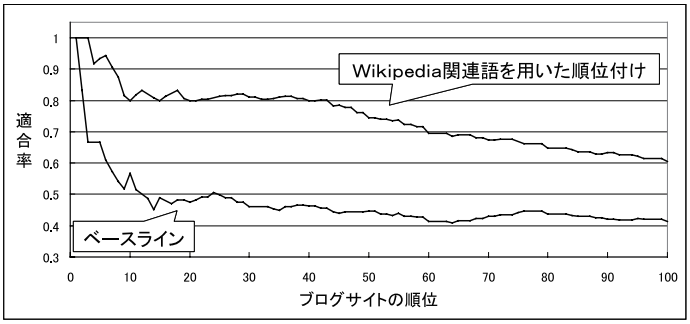
$$PostScore(p) = \sum_{t}(weight(type(t)) \times freq(t))$$

ここで，$weight(type(t))$ 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．関連語 t の種類 $type(t)$ としては，Wikipedia エントリタイトル，リダイレクト，エントリ本文中の太字，本文中における他エントリへのリンクのアンカーテキストの 4 種類を考慮し，それぞれの重みは 1 または 0 とする．評価実験を通して最適な重みの組み合わせを求め，全ての重みを 1 とした[*12]．さらに，ブログサイト s のスコアとして以下の式を用いる．

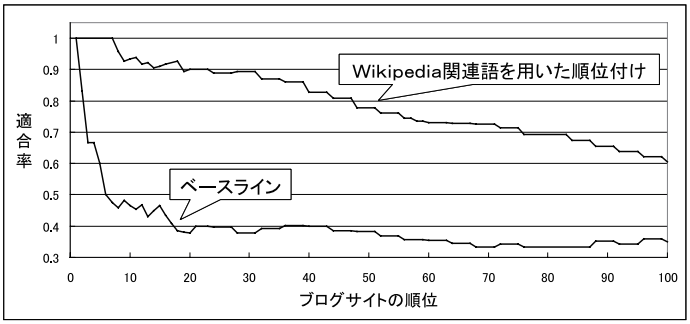
$$SiteScore(s) = \sum_{p} PostScore(p)$$

ただし，ブログ記事 p は，ブログサイト s に含まれるブログ記事である．

*12 これらの重みの設定の際には，まず，日本語 Wikipedia エントリに限定して 60 個のエントリを選定し，これを用いて収集したブログサイトを評価対象として，重みの調整を行った．具体的には，まず，各エントリタイトルをトピックとして，前節までの手順により，順位付け対象となるブログサイトを収集した．次に，その中から評価対象として用いるブログサイトをサンプリングして選定し，人手で「当該トピックについて詳細な記述を含むか否か」の判定を付与した．そして,「当該トピックについて詳細な記述を含む」ブログサイトをより上位に順位付けするように，Wikipedia 関連語の重みを設定した．さらに，4·4 節においてブログサイト検索手法の評価対象として用いた 3 トピック (日本語および英語) に対しても，Wikipedia 関連語の重みが最適な値に設定されていることを確認した．



(a) 日本語



(b) 英語

図 3 特定トピックのブログサイト検索における適合率の評価

4·4 評　　価

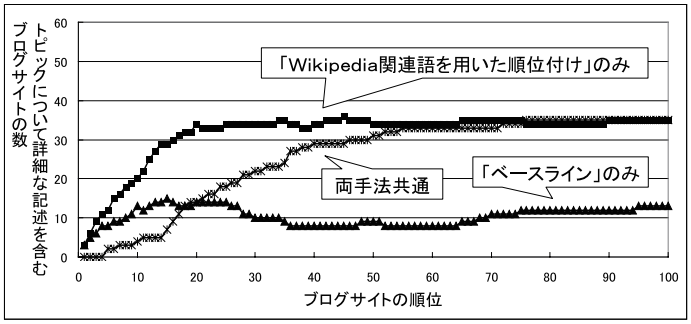
2 章で挙げた 7 トピックのうち，特に,「アルコール依存症 (alcoholism)」,「リストラ (restructuring)」,「著作権侵害 (copyright infringement)」の 3 トピックを対象として，4·1 節で述べた手法を用いて収集したブログサイトを，4·3 節で述べた手法で順位付けした結果の人手評価を行った．各トピックを対象として順位付けされた日英のブログサイトのうち，それぞれ，上位 20 ブログサイト，および，100 位までの 20 ブログサイトを等間隔にサンプリングした合計 40 ブログサイトを手動で評価した．評価結果としては,「当該トピックについて詳細な記述を含むか否か」の判定を付与した．横軸に各ブログサイトの順位をとり，適合率を縦軸にとってその推移をプロットしたものを図 3 の「Wikipedia 関連語を用いた順位付け」(提案手法) に示す．また，比較対象として，検索エンジン API によるブログ記事の順位付けを変更せずに，ブログサイト単位にまとめた順位付けに対しても，同様に合計 40 ブログサイトを手動で評価し，適合率を縦軸にとってその推移をプロットしたものを図 3 の「ベースライン」に示す．ただし，このプロットには，3 トピック分を平均した結果を示す．この結果から分かるように，日英どちらの言語においても，提案手法によりベースラインの適合率を大幅に改善することが分かる．

また，図 4 には，横軸に各ブログサイトの順位をとり,「トピックについて詳細な記述を含むブログサイトの数」(3 トピック分) の推移を，両手法の間で比較した．具体的には，各順位までで，両手法によって共通に出力されたブログサイト数，片方の手法によってのみ出力されたブログサイト数 (提案手法およびベースラインを区別して 1 プロットずつ) の比較を行った*13．この結果から明らかなように，提案手法のみによって出力されたブログサイト数は，ベースラインのみによって出力されたブログサイト数よりもはるかに多い．これにより，提案手法は,「トピックについて詳細な記述を含むブログサイト」のうち，既存の検索エンジン API において下位に順位付けされたサイトを上位に押し上げていることが分かる．

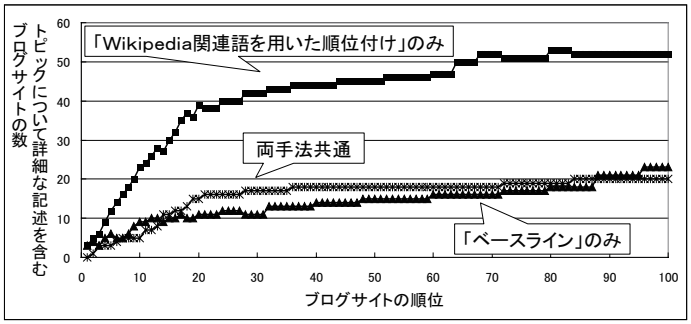
5. 日英ブログの言語対照分析

最後に，図 1 の枠組みにしたがって，上位に順位付けされたブログサイトおよびブログ記事を人手で分析し，さらに，ブログ中の記述を日英二言語間で比較・対照分析した結果の要約・抜粋を表 3 および表 4 に示す．ここで，各ブロガーの関心・意見の動向は，各ブロガーの国籍・居住地およびその文化的背景に依存するものであり，日本語・英語といった言語に依存するものではない．また，英語ブロガーの国籍・居住地・文化的背景について

*13 ここで,「両手法共通」のブログサイト数は，順位が下位になるにつれて単調増加する．一方，順位が上位の時点では，片方の手法によってのみ出力されていたブログサイトも，より下位の順位においては，もう一方の手法によっても出力されるようになる場合がある．このような場合には，順位が下位になるにつれて，片方の手法によってのみ出力されたブログサイト数が減少する，ということが起こり得る．実際に，図 4 (a) においては,「ベースライン」のみによって出力されたブログサイト数が単調増加せず，中間の順位において一時減少する傾向を示した．



(a) 日本語



(b) 英語

図 4 特定トピックのブログサイト検索における「トピックについて詳細な記述を含むブログサイトの数」の比較

は，米国等の英語圏の国を中心として多様な分布をしている可能性もあり得る．したがって，厳密には，各ブログの分析において，ブログ中の記述から各ブロガーの国籍・居住地・文化的背景を正確に同定したうえで，国間や文化間のブログ対照分析を行うことが望ましい．しかし，各ブログの分析において，ブログ中の記述から各ブロガーの国籍・居住地・文化的背景を正確に同定することは容易でない場合も予想されるため，本論文では，各ブロガーの国籍・居住地・文化的背景を考慮した分析は行わず，あくまで，日本語および英語の二言語間でのブログ対照分析を行う．ただし，以下で分析結果を示す各トピックについては，各英語ブロガーは，いくつかの例外を除いて，米国籍もしくは米国在住であると推測される．

7 トピックのうち，特に,「捕鯨」に関しては，日本語ブログと英語ブログの間に対極的な差異が観測された．上位に順位付けされた日本語ブロガーの多くは，海外の反捕鯨団体に反感を持っているのに対して，上位に順位付けされた英語ブロガーの多くは，捕鯨反対派であり，日本の捕鯨に対して批判的であった．

また,「臓器移植」に関しては，日本語ブロガーの間では，法的制度の整備に対する関心が高く，一方，英語ブロガーの間では，ドナー登録推奨に対する関心が高い．上位に順位付けされた英語ブロガーの多くは，米国籍もしくは米国在住であると推測され，社会背景的に，臓器移植制度の整備が遅れている日本と制度の整備が進んでいる米国との差異を反映する形で，上位に順位付けされたブロガーの関心・意見の違いが観測されたものと考えられる．

同様に,「アルコール依存症」,「リストラ」については，社会背景の影響の有無は不明であるが，上位に順位付けされたブロガーの関心事項の差異が観測された.「アルコール依存症」に関しては，日本語では，アルコール依存症患者自身によるブログが多く観測されたの対して，英語では，専門家等の第三者がアルコール依存症対策や治療法を紹介している事例が多かった.「リストラ」についても，日本語では，被リストラ体験者によるブログが比較的多く観測されたのに対して，英語では，経済学者やリストラする側の経営者等によるブログが多く観測された．

「サブプライムローン」については，日英いずれにおいても，上位に順位付けされたブロガーの関心事項の発端となる事象が米国発の金融危機である，という点は共通している．しかし，英語ブロガーの多くは米国籍もしくは米国在住であると推測され，米国での事態に対して関心を持っているのに対して，日本語ブロガーの多くは，それらの危機が日本経済に波及した結果生じた事態に対して関心を持っており，主要な関心事項の間に差異が観測される．

「喫煙」,「著作権侵害」についても，上位に順位付けされたブロガーの関心事項について，多少の言語間差異は認められるが，他のトピックほどは大きくない.「喫煙」に関しては，日英とも喫煙反対派のブロガーが多い．ただし，日本語ブロガーの中には，一部，喫煙者自身が喫煙賛成の意見を記述している事例が観測された.「著作権

表 3 評価用トピック (アルコール依存症，リストラ，著作権侵害) に関する日英ブログ中の意見の要約

評価用トピック — 概要	
(日本語ブログ中の意見)	(英語ブログ中の意見)
アルコール依存症 — ブログ空間においてアルコール依存症の人が多く観察される．	
アルコール依存症患者によるブログが多く，断酒日記，アルコール依存症対策・治療法について書いている。ブロガー例 ··· (1) アルコール依存症患者．(2) アルコール依存症経験者，10 年以上禁酒を継続．	アルコール依存症患者以外の第三者がアルコール依存症対策や治療法を紹介している事例が多い．アルコール依存症患者も存在するが，日本語よりは少ない．ブロガー例 ··· (1) アルコール依存症経験者．アルコール依存症患者からの相談と回答を紹介．(2) アルコール依存症に関する論文・ニュース・レポートを紹介．
リストラ — 金融危機による不況で多くの会社がリストラを行っている．	
被リストラ体験者によるブログが一定数存在する．リストラ後の失業保険や対策などを記述している．リストラに関するニュースを紹介しているブロガーが多い．ブロガー例 ··· (1) 被リストラ体験者．自分をリストラした会社を非難．(2) 被リストラ体験者の妻．リストラ後の節約生活を記述．	経済学など学術的な観点から企業のリストラを分析しているブロガーが多い．被リストラ体験者自身によるブログは観測されていない．ブロガー例 ··· (1) 経済・マーケティングに精通したブロガー．企業のリストラに関するレポートを多数紹介．(2) 株取引ブロガー．株価に影響する企業のリストラ情報を多く提供．
著作権侵害 — 著作権侵害に関する裁判が多数起きており，インターネット上ではファイル共有ソフトによる著作権侵害も問題となっている．	
日本における著作権侵害の裁判を紹介しているブロガーが多い．ファイル共有ソフトによる著作権侵害を批判しているブロガーがいる．ブロガー例 ··· (1) 著作権侵害に関する裁判に詳しく，どの判事が著作権侵害について厳しいか等を記述．(2) ファイル共有ソフトや音楽無料ダウンロードサイトによって，著作権に対する意識が薄れていると主張．	米国や欧州における著作権侵害の裁判を紹介しているブロガーが多い．音楽関係の著作権侵害に着目しているブロガーがいる．ブロガー例 ··· (1) 法律関係のブロガー．著作権侵害に関する本を出版．(2) 著作権に関する法律の改正を訴えるブロガー．音楽の著作権を独占しようとしている全米レコード協会を批判．

侵害」については，日本での裁判事例を紹介している日本語ブロガー，および，米国，欧州等での裁判事例を紹介している英語ブロガーが多く観測された．

6. 関 連 研 究

本論文の研究についての関連研究は，大きく分けると，ブログサイトの検索に関する研究，および，多言語情報源を対象とした情報分析の研究に分けることができる．

前者についての関連研究として，TREC の 2007 年度の Blog Distillation タスク [Macdonald 07] においては，ある特定のトピックについて検索したときに，そのトピックについて詳しく書かれていて，繰り返し見たいと思うブログサイトを検索するというタスクを行っている．このタスクにおいて上位の成績を収めた [Elsas 07] においては，本論文の手法と同様，Wikipedia エントリ中の他エントリへのリンクを用いた検索質問拡張が採用されている．一方，本論文では，Wikipedia 中のリダイレクトおよび太字といった，エントリとの関連性がより高い知識も合わせて用いてブログサイトの検索を行っている．また，ベースラインとして，既存の検索エンジン API によるブログサイトの順位付けからの改善を実証している．さらに，本論文では，日英二言語において提案手法の有効性を実証した．

その他には，ブロガーの熟知度に基づき，ブログサイトをランキングする手法 [中島 08] などがある．この手法では，マニアの多そうなキーワードを集めたマニア辞書をあらかじめ作成しておき，このマニア辞書に基づいてブログサイトの順位付けを行う．本論文の手法が，Wikipedia を知識源として利用するのに対して，この手法では，あらかじめマニア辞書を作成する必要がある点が大きく異なっている．

また，筆者らによる [川場 09] では，Wikipedia エントリによって指定されたトピックとブログサイトとの間の記述内容の対応を判定するタスクにおいて，機械学習によって対応の有無の判定を行っている．この研究では，記述内容の対応の有無を二値で判定することに焦点が当られているのに対して，本論文の手法では，Wikipedia エントリとブログサイトの間の記述内容の対応の度合いを測定することに焦点が当てられている．実際に，筆者らが，i) 記述内容の対応の有無の二値判定，ii) 記述内容の対応の度合いの測定，の二種類のタスクにおいて両者の手法の性能を比較したところ，タスク i) では [川場 09] の機械学習を用いた手法がやや上回り，タスク ii) では本

表 4 評価用トピック (捕鯨，臓器移植，喫煙，サブプライムローン) に関する日英ブログ中の意見の要約

評価用トピック — 概要	
(日本語ブログ中の意見)	(英語ブログ中の意見)
捕鯨 — 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．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．ブロガー例 ··· (1) 捕鯨賛成．反捕鯨団体を批判．(2) 捕鯨賛成．米国在住 12 年のブロガー．	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．ブロガー例 ··· (1) 捕鯨に関して中立的立場．日本に 30 年以上在住しているブロガー．(2) 捕鯨反対．動物愛護運動家のブロガー．(3) 捕鯨反対．シーシェパード派，反グリーンピース派．
臓器移植 — 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．ブロガー例 ··· (1) 病気腎移植のニュースを取り上げている．病気腎移植への反対意見を批判．(2) 脳死移植に反対．臓器移植法の改正は慎重に行うべきと主張．(3) 病気腎移植に反対．患者が完治するとは思えないと主張．	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．ブロガー例 ··· (1) 中国の違法臓器摘出を批判しているニュースを紹介．(2) 臓器提供に関するニュース記事を紹介．ドナー登録することを強く推奨．
喫煙 — 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．ブロガー例 ··· (1) 喫煙者と非喫煙者の間で対立が起きていることを指摘．(2) 喫煙反対．喫煙者はもっと喫煙マナーを守るべきと主張．(3) 喫煙は認知症の発症率を上げる可能性があると警告．	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．ブロガー例 ··· (1) 米国北部と米国南部の喫煙率を比較．また，タバコは米国の主要な農産物の一つだと主張．(2) 禁煙を強く推奨．喫煙は人体に悪影響を及ぼすだけであると主張．(3) 喫煙しないことが最も肺がんになりにくい方法であると主張．
サブプライムローン — 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．ブロガー例 ··· (1) 日本の大学の経営学科教授のブログ．日本市場はサブプライム問題に対して迅速な対応ができなかったことを指摘．(2) 日本の経済アナリストのブログ．誰も不動産の価格が下落するとは思っていなかったため，サブプライム問題の影響がより拡大したと指摘．(3) 近年の金融危機やサブプライム問題を引き起こした連邦準備制度理事会を批判．	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．ブロガー例 ··· (1) 連邦準備銀行は当初，サブプライム問題を深刻な問題として受け止めていなかったことを指摘．(2) サブプライムローン利用者は対策のしようがなかった．貸手側に大きな責任があると指摘．(3) いつか住宅バブルは弾けるとわかっていながらも，住宅バブルの影響でサブプライムローンを利用して家を購入した人が増加したことを指摘．

論文の手法がやや上回る，という結果であった．

一方，後者についての関連研究として，複数情報源からのニュースの多言語間差異分析を行っている研究 [Yangarber 07, Pouliquen 07, Yoshioka 08, Bautin 08] が挙げられる．[Yangarber 07] は，32 言語における 1000 以上の情報源を分析し伝染病に関するレポートをまとめあげる研究を行っている．[Pouliquen 07] では，32 言語におけるニュース記事群から特定の人物名を収集し，その人物の人間関係やその人物について言及している各国のニュース記事を継続的に分析する研究を行っている．[Yoshioka 08] は，複数の国の代表的なメディアが発信するニュースを情報源として，同一事象に対する各国のニュースの伝え方の差異分析をテーマとしている．[Bautin 08] では，9 言語間における同一事象に対する主観情報の差異分析の研究を行っている．これらの関連研究は主にニュース記事を対象に分析を行っている点で本論文とは異なる．

7. まとめと今後の課題

本論文では，近未来チャレンジ「Wikipedia マイニング」 [中山 09] の一環として，Wikipedia マイニングによって抽出した連想関係が情報検索タスクにおいて有用であること，および，Wikipedia に含まれる多言語知識を用いることにより，異言語横断文書検索が容易に実現でき，さらに，検索結果を対象として言語間差異の対照分析が実現できることを実証した．具体的には，本論文では，個々のブログ記事ではなく，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に注目し，そのような専門的内容を含むブログサイトを選択的に検索する手法を提案した．本論文の手法では，特

定のトピックについての詳細な情報を含むブログサイトを選択的に検索することを実現するために，各トピックについての Wikipedia エントリ中の記述を知識源として利用した．実際に，本論文の手法を既存の検索エンジン API と比較し，ブログサイトの検索性能において本論文の手法が優れていることを示した．さらに，本論文では，日英両言語を対象として，詳細な内容を記述したブログサイトが一定数存在するトピックを対象として，上述の手法により検索したブログサイト，および，各ブログサイトのブログ記事の内容を，日英二言語の間で比較・対照分析した．そして，特に，社会現象および社会問題に関するトピックについて，各トピックごとに興味深い言語間差異を観測した結果を示した．

今後は，日英二言語の間で，言語間差異の有無の自動判定に取り組むことが重要な課題の一つである．この課題においては，多言語での主観情報抽出技術 (例えば，[Evans 07, Wiebe 05]) を導入し，トピック「捕鯨」の場合のように，二言語間で同一の関心事項に対して意見の極性が反転していることを検出する必要がある．また，筆者らは，[中崎 09] において,「犯罪」分野を事例として，各トピックについて詳細な内容を記述為ているブログサイトに対して，日英両言語に共通の類型化が可能であることを示している．今後は，新規の分野のトピックに対しても，このような類型化の自動化手法を確立し，類型化結果の差異に基づいて言語間差異の有無の自動判定を実現することが期待される．

◇ 参 考 文 献 ◇

[Bautin 08] Bautin, M., Vijayarenu, L., and Skiena, S.: International Sentiment Analysis for News and Blogs, in *Proc. ICWSM*, pp. 19–26 (2008)

[Elsas 07] Elsas, J., Arguello, J., Callan, J., and Carbonell, J.: Retrieval and Feedback Models for Blog Distillation, in *Proc. TREC-2007 (Notebook)*, pp. 170–175 (2007)

[Evans 07] Evans, D. K., Ku, L.-W., Seki, Y., Chen, H.-H., and Kando, N.: Opinion Analysis across Languages: An Overview of and Observations from the NTCIR6 Opinion Analysis Pilot Task, in *Proc. 3rd Inter. Cross-Language Information Processing Workshop (CLIP2007)*, pp. 456–463 (2007)

[川場 09] 川場 真理子, 中崎 寛之, 横本 大輔, 宇津呂 武仁, 福原 知宏：Wikipedia 概念体系とブログ空間の間のトピック対応の推定, 日本データベース学会論文誌, Vol. 8, No. 1, pp. 17–22 (2009)

[Macdonald 07] Macdonald, C., Ounis, I., and Soboroff, I.: Overview of the TREC-2007 Blog Track, in *Proc. TREC-2007 (Notebook)*, pp. 31–43 (2007)

[中島 08] 中島 伸介, 稲垣 陽一, 草野 奉章：ブロガーの熟知度に基づいたブログランキング方式の提案, 電子情報通信学会第 19 回データ工学ワークショップ，第 6 回日本データベース学会年次大会 (DEWS2008) 論文集 (2008)

[中崎 09] 中崎 寛之, 阿部 佑亮, 宇津呂 武仁, 河田 容英, 福原 知宏, 神門 典子, 吉岡 真治, 中川 裕志, 清田 陽司：特定トピックの日英ブログ収集・分析・類型化: 事例研究, 情報処理学会研究報告, Vol. 2009, No. (2009–NL–194) (2009)

[中山 09] 中山 浩太郎, 伊藤 雅弘, ERDMANN, M., 白川 真澄, 道下 智之, 原 隆浩, 西尾 章治郎：Wikipedia マイニング: 近未来チャレンジキックオフ編, 人工知能学会論文誌, Vol. 24, No. 6, pp. 549–557 (2009)

[Pouliquen 07] Pouliquen, B., Steinberger, R., and Belyaeva, J.: Multilingual Multi-document Continuously-updated Social Networks, in *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp. 25–32 (2007)

[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)

[Yangarber 07] Yangarber, R., Best, C., von Etter, P., Fuart, F., Horby, D., and Steinberger, R.: Combining Information about Epidemic Threats from Multiple Sources, in *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp. 41–48 (2007)

[Yoshioka 08] Yoshioka, M.: IR Interface for Contrasting Multiple News Sites, in *Prof. 4th AIRS*, pp. 516–521 (2008)

〔担当委員：阿部 明典〕

2010 年 1 月 5 日　受理

著 者 紹 介

中崎　寛之

2010 年筑波大学大学院システム情報工学研究科博士前期課程修了．同年より株式会社 NTT データ勤務．ウェブ解析，自然言語処理の研究に従事．

川場　真理子

2009 年筑波大学大学院システム情報工学研究科博士前期課程修了．同年より日本電信電話株式会社 NTT サイバースペース研究所研究員．自然言語処理・情報検索の研究に従事．

横本　大輔

2010 年筑波大学第三学群工学システム学類卒業．現在 同大学大学院システム情報工学研究科博士前期課程在学中．ウェブ解析，自然言語処理の研究に従事．

宇津呂　武仁（正会員）

1989 年京都大学工学部 電気工学第二学科 卒業．1994 年同大学大学院工学研究科 博士課程電気工学第二専攻 修了．京都大学博士 (工学)．奈良先端科学技術大学院大学助手，豊橋技術科学大学講師，京都大学 講師を経て，2006 年より筑波大学 大学院システム情報工学研究科 知能機能システム専攻 助教授．2007 年より同准教授．自然言語処理の研究に従事．

福原　知宏（正会員）

2003 年 3 月奈良先端科学技術大学院大学情報科学研究科博士後期課程単位取得認定退学．博士 (情報工学)．郵政省通信総合研究所特別研究員，日本原子力研究所研究員，東京大学人工物工学研究センター特任助教を経て，2010 年より独立行政法人産業技術総合研究所サービス工学研究センター特別研究員．多言語ブログ記事を用いたテキストマイニング研究に従事．