Vol.2011-DBS-153 No.7
2011/11/3

Wikipediaを知識源とする 日英ブログ記事集合の観点分類と言語間対照分析

鈴木浩子[†1] 横本大輔[†1] 牧田健作[†1]
宇津呂武仁[†1] 河田容英[†2] 福原知宏[†3]

本論文では，同一のトピックについて，二つ以上の言語のブログにおいて関心を持たれている内容を言語間で対照分析する方式について研究を行う．本論文では特に，特定のトピックについて詳細な記述を含む英語ブログ記事に対して，日本語母語話者がその内容を理解する過程を支援することを目的として，同一の内容について記述した日本語ブログ記事を効率的に探索する枠組みについて述べる．本論文では，この枠組みを通して，日英ブログ空間における関心事項の言語間対照分析の一つの実現例を示す．

Facet Categorization of Japanese/English Blogs based on Wikipedia and their Comparative Analysis across Languages

Hiroko Suzuki,[†1] Daisuke Yokomoto,[†1]
Kensaku Makita,[†1] Takehito Utsuro,[†1]
Yasuhide Kawada[†2] and Tomohiro Fukuhara [†3]

This paper studies a framework of comparatively analyzing concerns within the blogospheres of more than one languages. Especially, given an English blog post concerning a specific topic, this paper proposes a framework of efficiently searching for Japanese blog posts, which include certain content almost the same as a passage in the English blog post. This framework aims at assisting the process of a Japanese native speaker's understanding the specific content of the English blog post. The proposed framework can be regarded as an example case of realizing comparative analysis of concerns across Japanese and English blogosphers.

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．その結果，多言語のブログ記事に対して，ある特定のトピックについて検索を行うことで，そのトピックが世界の各地域でどのように関心を持たれているのかを知ることができるようになった．

ここで，本論文では，同一のトピックについて，二つ以上の言語のブログにおいて関心を持たれている内容を言語間で対照分析する方式について研究を行う．本論文では特に，特定のトピックについて詳細な記述を含む英語ブログ記事に対して，日本語母語話者がその内容を理解する過程を支援することを目的として，同一の内容について記述した日本語ブログ記事を効率的に探索する枠組みについて述べる．本論文では，この枠組みを通して，日英ブログ空間における関心事項の言語間対照分析の一つの実現例を示す．

本論文で提案する「同一内容に関する日英ブログ記事の発見支援の枠組み」を図 1 に示す．この枠組みの概要を以下で述べる．

まず，特定のトピックについて，日本語および英語のブログ空間において，どのような観点のもとでブログ記事が書かれているかについての観点分布を提示する．図 1 においては，日本語トピック「地球温暖化」，および，英語トピック *"global warming"* について，「1. 日英ブログ記事集合の観点分布の作成」の段階において，「日本語ブログにおける観点集合」および「英語ブログにおける観点集合」がベン図の形式で示されている．この観点分布のベン図は，日本語トピック「地球温暖化」，および，英語トピック *"global warming"* に関して，日本語母語話者が観点の分布を俯瞰するために提示されるもので，技術的には，Wikipedia エントリ中の記述に基づいてブログ記事集合の観点分類を行う手法[3),7),8)] を用いることにより，観点の分布を作成する．この二言語観点分布を参照することにより，日本語母語話者は，興味を持った英語観点およびその観点についての英語ブログ記事を容易に指定すること

†1 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
†2 (株) ナビックス Navix Co., Ltd.
†3 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

が可能となる．

次に，本論文の枠組みにおいては，日本語母語話者が指定した英語ブログ記事に対して，同一の内容について記述した日本語ブログ記事を効率的に探索する過程を実現する．この枠組みにおいては，図 1 の「5. 相手言語 (日本語) ブログ記事の収集および順位付け」，および,「6. 同一内容に関する相手言語 (日本語) ブログ記事の発見支援」の過程に示すように，英語ブログ記事中から特徴的な英語キーワード (本論文の枠組みにおいては，英語 Wikipedia におけるエントリタイトル) を選定し，それを日本語訳としたものを検索クエリとして用いて，日本語ブログ記事を収集する．そして，それらの日本語ブログ記事の中から，英語ブログ記事と同一の内容を含むものを探索する．

本論文の枠組みにおいては，以上の手順を経ることにより，日本語母語話者が興味を持った英語ブログ記事の内容を裏付ける日本語ブログ記事を，比較的容易に探索する過程を支援することを実現している．

2. 同一内容に関する日英ブログ記事の発見支援の枠組み

本節では，図 1 に示す「同一内容に関する日英ブログ記事の発見支援の枠組み」の具体例を通して，本論文で提案する枠組みの概要を示す．

2.1 日英ブログ記事集合の観点分布の作成

このステップでは，前節で述べたように，日本語トピック「地球温暖化」，および，英語トピック *“global warming”* について,「日本語ブログにおける観点集合」および「英語ブログにおける観点集合」がベン図の形式で示される．日本語母語話者は，このベン図を参照することにより，日本語ブログ特有の観点，英語ブログ特有の観点，日本語ブログ・英語ブログ共通の観点，といった観点の分類を容易に俯瞰することができる．

2.2 分析対象観点の指定

次に，日本語母語話者は，前節で述べた観点分布に対して，その中の一つを分析対象観点として指定する．図 1 の例では，日本語ブログにおける出現頻度が相対的に小さく，英語特有の観点としてベン図に掲載されている観点 *“Inuit”*(イヌイット) を指定している．

2.3 ブログ記事の収集および順位付け

前節のステップにおいて指定された英語観点を利用して，英語ブログ記事を収集する．このステップにおいては，初期段階での英語トピック (英語初期トピックと呼ぶ．図 1 の例では，*“global warming”*) と，前節で指定した英語観点 (図 1 の例では，*“Inuit”*) の AND 検索により，英語ブログ記事を収集する．

2.4 分析対象ブログ記事の選定

日本語母語話者は，前節のステップにおいて収集されたブログ記事集合の中から，分析対象となる英語ブログ記事を選定する．ここでの英語ブログ記事選定の基準としては，

- 英語ブログ記事の内容が難解なため，以降のステップにおいて同一の内容に関する日本語ブログ記事を探索し，英語ブログ記事の内容理解を支援する．
- 英語ブログ記事の内容は理解できるが，その内容の信憑性に確証が持てないため，以降のステップにおいて同一の内容に関する日本語ブログ記事を探索し，日本語ブログ空間における裏付けがとれるか否かを検証する．
- 英語ブログ記事の内容に強い関心があるため，以降のステップにおいて同一の内容に関する日本語ブログ記事を探索し，日本語ブログ空間における同様の話題についての動向を把握する．

といったことが想定される．

2.5 相手言語 (日本語) ブログ記事の収集および順位付け

前節のステップにおいて選定された英語ブログ記事に対して，同一の内容に関する日本語ブログ記事の候補を収集するための日本語クエリを作成し，日本語ブログ記事を収集する．

具体的には，まず，英語初期トピックを Wikipedia の言語間リンクにより日本語エントリタイトルとしたもの (日本語初期トピックと呼ぶ．図 1 の例では,「地球温暖化」)，2.2 節で指定した英語観点を Wikipedia の言語間リンクにより日本語エントリタイトルとしたもの (図 1 の例では,「イヌイット」) の AND 検索により日本語ブログ記事を収集したものを，探索対象の日本語ブログ記事とする．また，これに加えて，英語ブログ記事中から英語キーワードの候補 (本論文の枠組みにおいては，英語 Wikipedia におけるエントリタイトル) を抽出し，Wikipedia の言語間リンクを用いて日本語エントリタイトルとしたもの (図 1 の例では，例えば,「人権」) を加えた三項組の AND 検索 (図 1 の例では，地球温暖化 AND イヌイット AND 人権) により日本語ブログ記事を収集したものも，同様に探索対象とする．

2.6 同一内容に関する相手言語 (日本語) ブログ記事の発見支援

日本語母語話者は，前節のステップにより収集されたブログ記事集合の一部を探索対象として選定する．具体的には，図 1 の例では，日本語初期トピック「地球温暖化」，および，英語観点「イヌイット」の AND 検索に収集されたブログ記事集合，さらに，これに,「人権」を加えた三項組の AND 検索により収集されたブログ記事集合，等の中から，一群のブログ記事集合を指定し，2.4 節で指定した英語ブログ記事と同一内容の日本語ブログ記事を探索する．そして，結果として，同一の内容について言及している日本語ブログ記事を発見

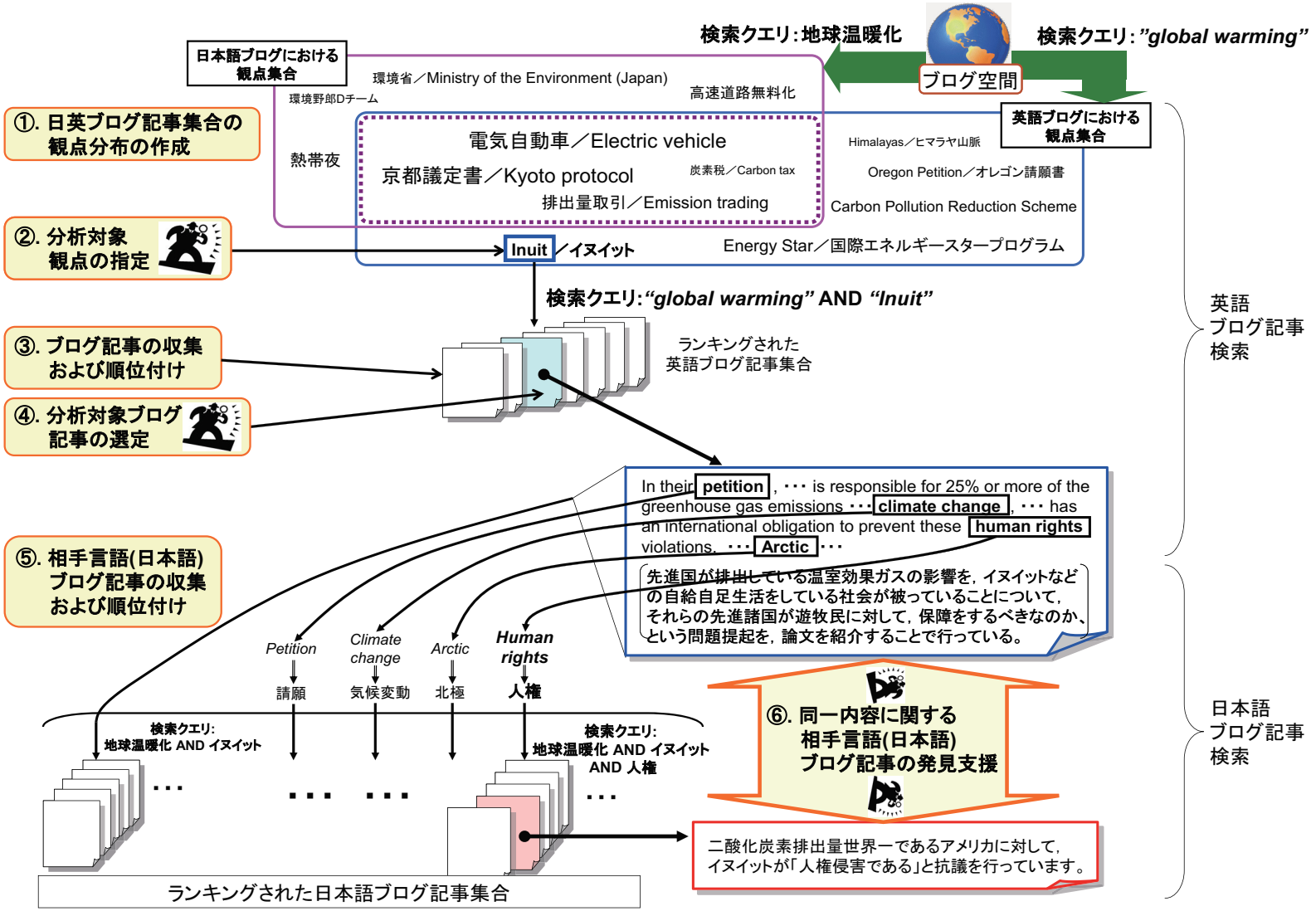

図 1 同一内容に関する日英ブログ記事の発見支援の枠組み

している．

3. 特定トピックの日英ブログ記事集合の観点分布の作成

本節では，Wikipedia エントリ中の記述に基づいてブログ記事集合の観点分類を行う手法[3),7),8)] を用いることにより，特定のトピックについて詳細な記述が含まれる日英ブログ記事集合における観点の分布を，ベン図の形式で俯瞰的に提示する手順について述べる．

以下では，観点分布作成の対象となる特定トピックのことを初期トピックと呼び，特に，そのうちの日本語でのトピック名を日本語初期トピック t_j^0，英語でのトピック名を英語初期トピック t_e^0 と記述する．さらに，t_j^0 および t_e^0 は，それぞれ，日本語 Wikipedia，および，英語 Wikipedia 中のエントリ名として登録されており，両者の間には，片方向もしくは両方向の言語間リンクが存在すると仮定する．

3.1 観点の収集

3.1.1 日本語観点の収集

日本語初期トピック t_j^0 をクエリとし，検索エンジン API として Yahoo! Search BOSS API[⋆1] を利用し，大手ブログホスト 8 社[⋆2] を指定してブログ記事の検索を行い，日本語ブログ記事集合 $D_j(t_j^0)$ を作成する．

次に，日本語初期トピック t_j^0 に対して，収集したブログ記事に付与する観点の集合 $F(t_j^0)$ を作成する．具体的には，まず，本文中に，日本語初期トピック t_j^0 が出現する日本語 Wikipedia エントリを f_j^0 とする．そして，f_j^0 のうち，ブログ記事集合 $D_j(t_j^0)$ において，エントリタイトル $t(f_j^0)$ の文書頻度が 30 以上となるものを選定し，観点集合 $F(t_j^0)$ を構成する．

$$F(t_j^0) = \left\{ f_j^0 \;\middle|\; \mathrm{df}(\ D_j(t_j^0),\ t(f_j^0)\ ) \ \geq 30 \right\}$$

3.1.2 英語観点の収集

日本語の場合と同様に，英語初期トピック t_e^0 をクエリとし，検索エンジン API として Yahoo! Search BOSS API を利用し，大手ブログホスト 4 社[⋆3] を指定してブログ記事の検索を行い，英語ブログ記事集合 $D_e(t_e^0)$ を作成する．

次に，英語初期トピック t_e^0 に対して，収集したブログ記事に付与する観点の集合 $F(t_e^0)$ を作成する．日本語の場合と同様に，本文中に，英語初期トピック t_e^0 が出現する英語 Wikipedia エントリを f_e^0 とする．そして，f_e^0 のうち，ブログ記事集合 $D_e(t_e^0)$ において，エントリタイトル $t(f_e^0)$ の文書頻度が 11 以上となるものを選定し，観点集合 $F(t_e^0)$ を構成する．

$$F(t_e^0) = \left\{ f_e^0 \;\middle|\; \mathrm{df}(\ D_e(t_e^0),\ t(f_e^0)\ ) \ \geq 11 \right\}$$

3.2 日英観点分布の作成

前節の手順により得られた日本語観点集合 $F(t_j^0)$ 中の各観点 f_j，および，英語観点集合 $F(t_e^0)$ の各観点 f_e について，Wikipedia の言語間リンクを用いることにより，日英対訳観点組 $\langle f_j, f_e \rangle$ を作成する．ただし，ここで，少なくとも，$f_j \in F(t_j^0)$ または $f_e \in F(t_e^0)$ のいずれか一方が成り立ち，f_j と f_e の間には片方向もしくは両方向の言語間リンクが存在することを必要条件とする．そして，この日英対訳観点組 $\langle f_j, f_e \rangle$ を集めた集合 $F\big(\langle t_j^0, t_e^0 \rangle\big)$ を作成する．ここで，この日英対訳観点組集合 $F\big(\langle t_j^0, t_e^0 \rangle\big)$ は，以下の三種類の部分集合に分割される．

日英共通観点集合 $F_{je}\big(\langle t_j^0, t_e^0 \rangle\big)$　日本語観点 f_j，および，英語観点 f_e の双方が，それぞれ日本語観点集合 $F(t_j^0)$，および，英語観点集合 $F(t_e^0)$ に含まれる．

日本語特有観点集合 $F_j\big(\langle t_j^0, t_e^0 \rangle\big)$　日本語観点 f_j のみが日本語観点集合 $F(t_j^0)$ に含まれ，英語観点 f_e は英語観点集合 $F(t_e^0)$ に含まれない．

英語特有観点集合 $F_e\big(\langle t_j^0, t_e^0 \rangle\big)$　英語観点 f_e のみが英語観点集合 $F(t_e^0)$ に含まれ，日本語観点 f_j は日本語観点集合 $F(t_j^0)$ に含まれない．

$$F\big(\langle t_j^0, t_e^0 \rangle\big) \ = \ F_{je}\big(\langle t_j^0, t_e^0 \rangle\big) \bigcup F_j\big(\langle t_j^0, t_e^0 \rangle\big) \bigcup F_e\big(\langle t_j^0, t_e^0 \rangle\big)$$

例えば，図 1 の「1. 日英ブログ記事集合の観点分布の作成」の部分においては，日本語初期トピック t_j^0 =「地球温暖化」，および，英語初期トピック t_e^0 = *"global warming"* の場合について，日英共通観点集合 $F_{je}\big(\langle t_j^0, t_e^0 \rangle\big)$，日本語特有観点集合 $F_j\big(\langle t_j^0, t_e^0 \rangle\big)$，および，英語特有観点集合 $F_e\big(\langle t_j^0, t_e^0 \rangle\big)$ をそれぞれ示す[⋆4]．

4. 特定の観点に関するブログ記事の収集

本節では，前節で作成した日英観点分布を参照して，利用者が特定の観点を指定し，指定された観点に関するブログ記事を収集する方式について述べる．

⋆1 http://developer.yahoo.com/search/boss/
⋆2 fc2.com，yahoo.co.jp，yaplog.jp，ameblo.jp，goo.ne.jp，livedoor.jp，Seesaa.net，hatena.ne.jp
⋆3 blogspot.com，wordpress.com，typepad.com，multiply.com

⋆4 ただし，図 1 においては，日本語観点 f_j が言語間リンクを持たない場合には，観点組 $\langle f_j, - \rangle$ を作成し，逆に，英語観点 f_e が言語間リンクを持たない場合には，観点組 $\langle -, f_e \rangle$ を作成している．

4.1 日英観点分布における分析対象観点の指定

利用者は，まず，前節で作成した日英共通観点集合 $F_{je}\big(\langle t_j^0, t_e^0\rangle\big)$，日本語特有観点集合 $F_j\big(\langle t_j^0, t_e^0\rangle\big)$，および，英語特有観点集合 $F_e\big(\langle t_j^0, t_e^0\rangle\big)$ のいずれかから，以降の分析の対象とする日本語観点 f_j^1 または英語観点 f_e^1 を指定する．

以下，本論文においては，特に英語観点 f_e^1 を指定した場合の分析手順について述べる．

4.2 ブログ記事の収集

次に，英語初期トピック t_e^0，および，前節で指定された英語観点 f_e^1 の両方に関連する英語ブログ記事を収集する．具体的には，検索エンジン API として Yahoo! Search BOSS API を利用し，3.1.2 節で述べた大手ブログホスト 4 社を指定して，英語初期トピック t_e^0 と英語観点 f_e^1 の二つの AND 検索により英語ブログ記事を収集し，英語ブログ記事集合 $D_e(t_e^0, f_e^1)$ を作成する．

4.3 ブログ記事の順位付け

前節で作成した英語ブログ記事集合 $D_e(t_e^0, f_e^1)$ 中の英語ブログ記事の順位付けを行い，上位のブログ記事から順に利用者に提示する．順位付けの方法としては，ブログ記事と Wikipedia エントリの類似度に基づく方法[7),8)]，および，検索エンジン API による順位付けをそのまま用いる方法の二通りが考えられる．

このうち，以下では，ブログ記事と Wikipedia エントリの類似度に基づく方法について述べる．

ブログ記事と Wikipedia エントリの類似度に基づく方法

この方法においては，ブログ記事 d と Wikipedia エントリ e の類似度[7),8)] $Sim(e,d)$ を用いて，類似度の降順にブログ記事を順位付けする．

具体的には，英語ブログ記事 d_e と英語初期トピック t_e^0 のエントリ $e(t_e^0)$ との類似度 $Sim(e(t_e^0), d_e)$，および，英語ブログ記事 d_e と観点 f_e^1 のエントリ $e(f_e^1)$ との類似度 $Sim(e(f_e^1), d_e)$ の和 $Sim(e(t_e^0), e(f_e^1), d_e)$ を算出し，英語ブログ記事の順位付けにおいてはこの類似度 $Sim(e(t_e^0), e(f_e^1), d_e)$ を用いる．

$$Sim(e(t_e^0), e(f_e^1), d_e) \;=\; Sim(e(t_e^0), d_e) + Sim(e(f_e^1), d_e)$$

ここで，ブログ記事 d と Wikipedia エントリ e の類似度 $Sim(e,d)$ を算出する際には，まず Wikipedia エントリ e の本文中に含まれる重要な語を関連語として抽出し，Wikipedia エントリ e を関連語の集合 $R(e)$ として表現する．そして，Wikipedia エントリ e の関連語 $r(\in R(e))$ を次元とするベクトル表現の内積により類似度を定義する．

4.4 分析対象ブログ記事の選定

4.2 節で作成された英語ブログ記事集合 $D_e(t_e^0, f_e^1)$ の中から，分析対象となる英語ブログ記事 d_e^1 を選定する．ただし，ここで，英語ブログ記事を選定する際の基準は，2.4 節で述べた通りである．

5. 同一内容に関する相手言語ブログ記事の発見支援

本節では，前節で選定した英語ブログ記事 d_e^1 に対して，同一の内容に関する記述を含む相手言語 (本論文では日本語) ブログ記事を発見する過程を支援する方式について述べる．

具体的には，まず，英語 Wikipedia の言語間リンクを用いることにより，英語初期トピック t_e^0，および，4.1 節で指定した分析対象英語観点 f_e^1 から，それぞれ，日本語初期トピック t_j^0，および，日本語観点 f_j^1 を得る．そして，検索エンジン API として Yahoo! Search BOSS API を利用し，3.1.1 節で述べた大手ブログホスト 8 社を指定して，日本語初期トピック t_j^0 と日本語観点 f_j^1 の二つの AND 検索により日本語ブログ記事を収集し，日本語ブログ記事集合 $D_j(t_j^0, f_j^1)$ を作成する．

次に，4.3 節で述べた手法と同様の手法により，日本語ブログ記事集合 $D_j(t_j^0, f_j^1)$ 中の日本語ブログ記事の順位付けを行い，上位のブログ記事から順に利用者に提示することにより，英語ブログ記事 d_e^1 の内容に関連する記述を含む日本語ブログ記事発見を支援する．

また，以上の手順により，英語ブログ記事 d_e^1 の内容に関連する記述を含む日本語ブログ記事が発見できない場合には，英語ブログ記事 d_e^1 の本文テキスト中から，重要な手がかりとなると予測される英語 Wikipedia エントリタイトル f_e^2 を指定する．

次に，英語 Wikipedia の言語間リンクを用いることにより，英語初期トピック t_e^0，4.1 節で指定した分析対象英語観点 f_e^1，および，上記の英語エントリタイトル f_e^2 から，それぞれ，日本語初期トピック t_j^0，日本語観点 f_j^1，および，日本語エントリタイトル f_j^2 を得る．そして，検索エンジン API として Yahoo! Search BOSS API を利用し，3.1.1 節で述べた大手ブログホスト 8 社のドメインを対象として，日本語初期トピック t_j^0，日本語観点 f_j^1，および，日本語エントリタイトル f_j^2 の三つの AND 検索により日本語ブログ記事を収集し，日本語ブログ記事集合 $D_j(t_j^0, f_j^1, f_j^2)$ を作成する．

この場合も，二つ組の AND 検索の場合と同様に，日本語ブログ記事集合 $D_j(t_j^0, f_j^1, f_j^2)$ 中の日本語ブログ記事の順位付けを行い，上位のブログ記事から順に利用者に提示することにより，英語ブログ記事 d_e^1 の内容に関連する記述を含む日本語ブログ記事発見を支援する．

例えば，図 1 の例においては，英語ブログ記事 d_e^1 に対して，同一の内容に関する記述を

含む日本語ブログ記事を発見する過程を支援するために，英語エントリタイトル f_e^2 として *"Human rights"*(人権) を選定している．そして，AND 検索「地球温暖化 AND イヌイット AND 人権」によって日本語ブログ記事を収集し，英語ブログ記事 d_e^1 の内容に関連する記述を含む日本語ブログ記事を効率よく発見している．

6. 分 析 例

本節では，初期トピックとして，

- ⟨ 地球温暖化, *"global warming"*⟩,
- ⟨ トヨタ・プリウス, *"Toyota Prius"*⟩

の二例を対象として分析を行った結果を示す．

6.1 検索エンジン API

本節では，分析において日英ブログ記事を収集する際に実際に使用した検索エンジン API，および，日英ブログ記事収集の時期について述べる．

6.1.1 日英観点の収集

3.1.1 節の日本語観点の収集においては，Yahoo! Japan API[⋆1]を利用し，2010 年 7 月上旬に，3.1.1 節で述べた大手ブログホスト 8 社のドメインを対象としてブログ記事の収集を行った．一方，3.1.2 節の英語観点の収集においては，Yahoo! Search BOSS を利用し，2010 年 12 月中旬に，3.1.2 節で述べた大手ブログホスト大手 4 社のドメインを対象としてブログ記事の収集を行った．

6.1.2 特定の観点に関するブログ記事の収集

4.2 節および 5 節における，AND 検索の検索クエリを用いた英語ブログおよび日本語ブログの収集は，2011 年 10 月上旬に行った．また，検索エンジン API の設定は，4.2 節および 5 節における説明のものをそのまま用いた．検索されたブログ記事の順位付けとしては，検索エンジン API による順位付けをそのまま用いた．

6.2 分 析 結 果

以上の設定のもとで分析を行った結果を表 1 に示す．表中には，「分析対象英語観点」，「分析対象の英語ブログ記事収集のための検索クエリ」，「分析対象の英語ブログ記事の要旨」，「英語ブログ記事に出現する Wikipedia エントリのタイトル」，「同一内容に関する日本語ブログ記事発見のための検索クエリ」，「同一内容に関する日本語ブログ記事の発見例」を，それぞれ順に示す．

6.2.1 初期トピック: ⟨ 地球温暖化, *"global warming"*⟩

初期トピックが ⟨ 地球温暖化, *"global warming"*⟩ の場合において，分析対象英語観点を，それぞれ，*"Inuit"*(イヌイット)，*"Himalayas"*(ヒマラヤ山脈)，*"Kyoto Protocol"*(京都議定書) とした場合の結果を表 1 (a) に示す．この場合，*"Inuit"*(イヌイット)，および，*"Himalayas"*(ヒマラヤ山脈) は英語特有観点であるのに対して，*"Kyoto Protocol"*(京都議定書) は日英共通観点である．

6.2.2 初期トピック: ⟨ トヨタ・プリウス, *"Toyota Prius"*⟩

初期トピックが ⟨ トヨタ・プリウス, *"Toyota Prius"*⟩ の場合において，分析対象英語観点を *"recall"*(リコール) とした場合の結果を表 1 (b) に示す．ここで，*"recall"*(リコール) は日英共通観点である．

この例の場合においては，分析対象英語観点として *"recall"*(リコール) を指定して，*"Prius"* AND *"recall"* を検索クエリとして英語ブログ記事収集を行ったところ，一連の報道に対して批判的な論調の英語ブログ記事を発見した．この英語ブログ記事には，多種多様な英語 Wikipedia エントリタイトルが含まれているが，その中でも特に特徴的なものとして，*"brand"*(ブランド) を指定して同一内容に関する日本語ブログ記事の探索を行うことにより，一連の報道に対して，日本の自動車業界を心配する論調の日本語ブログ記事を効率よく発見することができた．

7. 関 連 研 究

本研究に関する関連研究として，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[1),5),6),9)] が挙げられる．文献 6) 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．文献 5) 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．文献 9) 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．文献 1) 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．

一方，我々は，これまでに，文献 4) において，特定のトピックについての日英ブログ記事集合を収集し，その記述内容を日英二言語間で比較対照分析する方式を提案し，その有

⋆1 http://developer.yahoo.co.jp/webapi/search/websearch/

効性について評価を行った．文献 4) における成果と比較すると，本論文においては，一つトピックの全体に関連するブログ記事集合を分析対象とするのではなく，文献 2)，3)，7)，8) の手法により，特定のトピックについての観点分布を日英二言語で提示した結果に対して，特定の観点，および，その観点についての記述を含む英語ブログ記事を利用者に指定させる点，および，指定された特定の英語ブログ記事に焦点を当てて，日英二言語間での言語間対照分析を実現する点が大きく異なる．

8. おわりに

本論文では，同一のトピックについて，二つ以上の言語のブログにおいて関心を持たれている内容を言語間で対照分析する方式について述べた．本論文では特に，特定のトピックについて詳細な記述を含む英語ブログ記事に対して，日本語母語話者がその内容を理解する過程を支援することを目的として，同一の内容について記述した日本語ブログ記事を効率的に探索する枠組みを提案した．本論文では，この枠組みを通して，日英ブログ空間における関心事項の言語間対照分析の一つの実現例を示した．

参考文献

1) Bautin, M., Vijayarenu, L. and Skiena, S.: International Sentiment Analysis for News and Blogs, *Proc. ICWSM*, pp.19–26 (2008).
2) Lim, D., Yokomoto, D., Makita, K., Utsuro, T. and Fukuhara, T.: Utilizing Wikipedia as a Knowledge Source in Categorizing Topic related Korean Blogs into Facets, 言語処理学会第 17 回年次大会論文集，pp.876–879 (2011).
3) 牧田健作，横本大輔，鈴木浩子，宇津呂武仁，河田容英，福原知宏：Wikipedia を多言語知識源とするブログ集合の話題分析，電子情報通信学会技術研究報告, NLC2011-18, pp.95–100 (2011).
4) 中崎寛之，川場真理子，横本大輔，宇津呂武仁，福原知宏：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，人工知能学会論文誌， Vol.25, No.5, pp.613–622 (2010).
5) Pouliquen, B., Steinberger, R. and Belyaeva, J.: Multilingual Multi-document Continuously-updated Social Networks, *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp.25–32 (2007).
6) Yangarber, R., Best, C., von Etter, P., Fuart, F., Horby, D. and Steinberger, R.: Combining Information about Epidemic Threats from Multiple Sources, *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp.41–48 (2007).
7) Yokomoto, D., Makita, K., Utsuro, T., Kawada, Y. and Fukuhara, T.: Utilizing Wikipedia in Categorizing Topic related Blogs into Facets, *Proc. 12th PACLING,* #20 (2011).
8) 横本大輔，林　東權，牧田健作，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：特定トピックに関するブログ記事集合の観点分類における Wikipedia の利用，第 3 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集 (2011).
9) Yoshioka, M.: IR Interface for Contrasting Multiple News Sites, *Prof. 4th AIRS*, pp.516–521 (2008).

表 1 同一内容に関する相手言語ブログ記事発見手順の例

分析対象英語観点	分析対象の英語ブログ記事収集のための検索クエリ	分析対象の英語ブログ記事の要旨	英語ブログ記事に出現する Wikipedia エントリのタイトル	同一内容に関する日本語ブログ記事発見のための検索クエリ	同一内容に関する日本語ブログ記事の発見例
(a) 初期トピック: 〈 地球温暖化, *"global warming"*〉					
"Inuit" (イヌイット)	"global warming" AND "Inuit"	In their petition, ··· is responsible for 25% or more of the greenhouse gas emissions ··· climate change, ··· has an international obligation to prevent these human rights. (先進国が排出している温室効果ガスの影響を，イヌイットなどの自給自足生活をしている社会が被っていることについて，それらの先進諸国が遊牧民に対して，保障をするべきなのか、という問題提起を，論文を紹介することで行っている．)	"Human rights"(人権), "United States"(アメリカ合衆国), "Climate change"(気候変動), "Arctic"(北極), "China"(中国), "Developed"(先進国), "Greenhouse gas"(温室効果ガス), "Intergovernmental Panel on Climate Change"(気候変動に関する政府間パネル), "Kyoto Protocol"(京都議定書), "United Nations"(国際連合), ···	"地球温暖化" AND "イヌイット" AND "人権"	二酸化炭素排出量世界一であるアメリカに対して，イヌイットが「人権侵害である」と抗議を行っている．
"Himalayas" (ヒマラヤ山脈)	"global warming" AND "Himalayas"	The meltdown of glaciers due to global warming has sent a chill through the Himalayan region. Over the last couple of years, this mountainous country has recorded a hazy winter, hotter summer months, reduced rain fall and frequent landslides, which experts attribute to climatic change. (地球温暖化の影響で，ヒマラヤ周辺では，氷河が溶けたり，地すべりを起こすなど，異変が起こっている．)	"Glacier"(氷河), "Asia"(アジア), "Bhutan"(ブータン), "Deforestation"(森林破壊), "Everest"(エベレスト), "India"(インド), "Indus"(インダス川), "Nepal"(ネパール), "Tibetan Plateau"(チベット高原), "Western"(ウエスタン), ···	"地球温暖化" AND "ヒマラヤ" AND "氷河"	IPCC が，報告書の,「2035 年までにヒマラヤの氷河が消失」という記述が間違っていたことを表明したことに対して，IPCC の報告は信用できないと述べている．
"Kyoto Protocol" (京都議定書)	"global warming" AND "Kyoto Protocol"	As the massive global warming fraud implodes, the one aspect of it that has not been explored in depth is the equally massive waste of billions of dollars spent by the United States and nations around the world, we were told, to avoid global warming.（温暖化は科学的な根拠のない，大規模な金額の絡む詐欺だということを主張，政治課題として利用することを批判している.）	"Fraud"(詐欺), "Carbon dioxide"(二酸化炭素), "China"(中国), "Climate change"(気候変動), "Government"(政府), "Gross domestic product"(国内総生産), "Intergovernmental Panel on Climate Change"(気候変動に関する政府間パネル), "Senate"(元老院), "United Nations"(国際連合), "United States"(アメリカ合衆国), ···	"地球温暖化" AND "京都議定書" AND "詐欺"	京都議定書の排出取引がビジネス目的のものであることを述べ，温暖化が詐欺であることを主張している動画を紹介している．
(b) 初期トピック: 〈 トヨタ・プリウス, *"Toyota Prius"*〉					
"2009-2011 Toyota vehicle recalls" (トヨタ自動車の大規模リコール (2009 年-2010 年))	"Prius" AND "recall"	Nikkei business daily said in an editorial: "Words alone cannot settle the situation. Toyota represents Japan and its shaking could lead to a loss of trust for the entire Japan brad." (日本のメディアが，トヨタの問題が国内全体のブランドの信頼喪失につながるおそれがあると批判していることを取り上げている．)	"Brand"(ブランド), "Accident"(事故), "Complaint"(苦情), "Safety"(安全性), "Anti-lock braking system"(アンチロック・ブレーキ・システム), "Brake"(ブレーキ), "Class Action Lawsuit"(複雑訴訟形態), "Akio Toyoda"(豊田章男), "General Motors"(ゼネラルモーターズ), "Japan"(日本), ···	"プリウス" AND "リコール" AND "ブランド"	プリウスのリコールによって，ハイブリッドカーによって牽引されている，日本のクルマ業界の勢いが落ちてしまうことを心配している．