

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

福原知宏¹・宇津呂武仁²・中川裕志³

¹ 東京大学人工物工学研究センター 価値創成イニシアティブ(住友商事)寄附研究部門
(URL: <http://www.race.u-tokyo.ac.jp/~fukuhara/>)

² 筑波大学大学院 システム情報工学研究科 (URL: <http://nlp.iit.tsukuba.ac.jp/>)

³ 東京大学情報基盤センター 図書館電子化研究部門 (URL: <http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/>)

本論文では複数の言語で記述されたウェブログ記事を対象として言語間の語彙出現傾向比較を行う言語横断型関心解析システムを提案する。提案システムを用いた定量的分析と韓国語話者による定性分析の事例を報告する。

1. はじめに

今日、インターネット上では様々な国や地域においてウェブログ(ブログ)を用いた情報発信が盛んである。筆者らはブログ記事と新聞記事から社会の関心動向を解析するシステムの研究開発を行い、日本国内の社会問題に関する関心動向について調査を行ってきた¹⁾。一方、今日の社会問題は鳥インフルエンザや地球温暖化に代表されるように国や地域を越えて世界共通の問題として存在する。こうしたグローバルな問題の解決には、言語の壁を越えた関心動向の把握と分析が必要となる。

一方、自然言語処理の観点から言語横断的な関心動向分析の実現と実用化には、十分に高性能な機械翻訳技術の実現が不可欠である。しかし今日のブログ記事のように個人の主観情報や口語表現が多用される場合、現在の機械翻訳技術のレベルでは十分に内容を把握できる出力を期待するのは難しい。現実には様々な言語で記述されたブログ記事から有用な情報を得るには、多少なりともその言語が理解できる調査者の手を借りざるを得ない。ただしその際には、膨大な量のブログ記事から、真に有用な情報を含むと期待できる文書をいかにして効率良く選定し、翻訳者による翻訳の対象とするか、という点が最も重要な技術的課題である。

本論文では上の技術的課題を解決する言語横断型ウェブログ関心解析システムを提案し、その基本的な動作を概説する。また本システムにより、実際どの程度各国に特徴的な関心や情報を収集できるか推定するために、韓国語を母語とする調査者により各国語ブログ記事中の記述を定性的・定量的に分析した結果得られた特徴的な関心や情報の事例を紹介する。

本論文の構成は次の通りである。2.では言語横断型関心解析システムの目的と概要について述べ、3.では提案システムで得られた解析事例と韓国語を母語とする調査者へのインタビューから日韓の関心の相違についての定性分析について述べる。4.では本論文のまとめと今後の課題について述べる。

2. 言語横断型ウェブログ関心解析システム

本節では提案システムの目的と概要について述べる。

2.1 システムの目的

提案システムの目的は様々な言語で記述されたブログ記事の収集と解析を通じて世界の人々の関心動向を探ることにある。現在、ブログを対象とする多くの検索・解析システムが提案されている²⁾が、これらはいずれも単言語を対象としていた。これに対し本研究では様々な言語で記述された世界の人々のブログ記事の収集と解析を行い、ある話題について世界の人々の関心の比較を行うことで、国際対話のための情報基盤となることを目指す。

2.2 システムの概要

Fig.1 にシステムの概要を示す。提案システムは現在、日本語、中国語、韓国語、英語の4言語で記述されたブログサイトからRSS¹⁾ならびにAtomフィード²⁾を収集する。Table 1に収集記事数、登録サイト数、収集期間、1日あたりの平均収集記事数を示す。

収集したフィードからはdescriptionもしくはcontent要素の記事を見なし、キーワードを抽出し、記事を各言語のデータベースに格納する。キーワード抽出には形態素解析器とPOS Taggerを用いた。日本語にはJuman³⁾を、中国語にはICTCLAS³⁾を、韓国語にはKLT⁴⁾を、英語にはpostagger⁵⁾を用いた⁴⁾。

システムは利用者からの要求に応じ、記事検索や共起語検索、月間の話題語の抽出等を行う¹⁾。本システムでは検索語を複数の言語に翻訳して各言語の記事を検索することで各言語における関心を表す関心比較グラフ(Fig. 3に例を示す)を出力する。翻訳処理については2.3で述べる。

¹ <http://ja.wikipedia.org/wiki/RSS>

² <http://ja.wikipedia.org/wiki/Atom>

³ <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>

⁴ <http://nlp.kookmin.ac.kr/HAM/kor/>

⁵ <http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/postagger/>

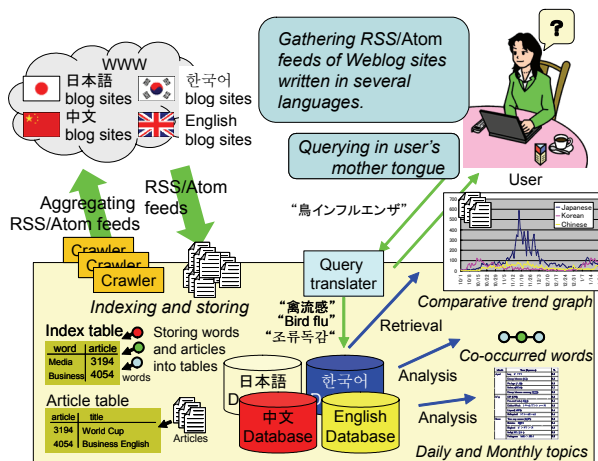


Fig. 1 システム概要

Table 1 記事数とサイト数 (2007年1月24日17時現在)

	日本語	中国語	韓国語	英語
記事数	129,218,107	6,647,082	25,707,257	3,121,588
サイト数	2,521,127	585,341	461,762	64,790
収集期間	1,042日	784日	541日	75日
1日の収集記事数	50万記事	2万記事	80万記事	7万記事



Fig. 2 Wikipediaを用いた対訳検索処理

2.3 対訳検索について

提案システムは検索語を他の言語に翻訳し、各言語のブログ記事における関心动向を検索する。この翻訳には Wikipedia⁶と辞書を用いる。Wikipediaは不特定多数の利用者によって自発的に管理されているため、既存の辞書に掲載されていない用語でも掲載されている場合があるため、これを利用することにした。

Wikipediaを用いた対訳表現検索処理の概要を Fig. 2 に示す。Wikipediaを用いた検索では、検索語を Wikipedia 上で検索し、検索結果の HTML ページに他の言語の同一エントリへのリンクがあればこのリンクを辿り、目標言語での表現を得ることで訳語を得る。他の言語への同一エントリへのリンクが含まれていない場合は処理を終了する。なお現在は日本語検索語を入力とするが、将来的には任意の言語での入力を目指す。

⁶ <http://ja.wikipedia.org/>

"クローン"に関する関心动向(記事数)

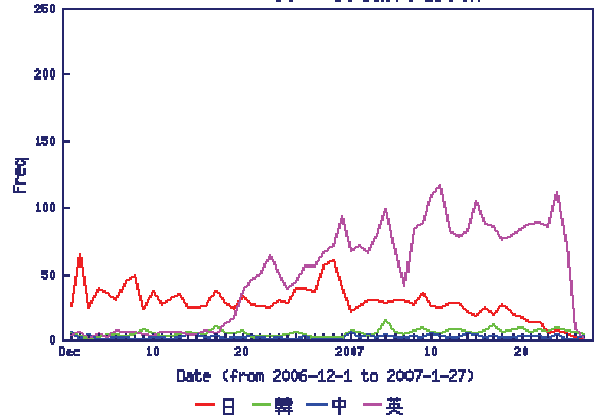


Fig. 3 “クローン”に関する言語横断検索 (記事数比較)

3. 分析事例

ここでは“クローン”に関する日韓での関心比較と、韓国語話者による定性的分析について述べる。

3.1 “クローン”に対する日韓関心比較

Fig. 3 に“クローン”の比較検索結果 (2006年12月1日から2007年1月27日まで) を示す⁷。記事数はそれぞれ1593(日)、80(中)、272(韓)、2847(英)であった。英語と日本語の件数が多く、韓国語、中国語の記事数は少ない。以下では日本語と韓国語の共起語から各言語における関心を比較する。

日本語圏における“クローン”への関心

日本語における共起語を Table 2 と Fig. 4 に示す。Table 2 はこの期間における共起語の上位5語である。クローン病とクローン携帯への関心が混ざった結果となった。Fig. 4 は共起語の時間軸上の推移である。“クローン”に対して日本語では定常的にクローン病(Crohn's disease)への関心(図中①)が見られたほか、クローン携帯(図中②)やクローン動物の食品利用への関心(図中③)が見られた。次に述べるがこれらの関心は韓国語圏とはまったく異なることが分かった。

韓国語圏における“クローン”への関心

韓国語における“クローン”の訳語として(1)音読みの“클론”(クローン)と(2)意識である“복제”(複製)について調査した。

第1に、音読みのクローンについて共起語上位5語(Table 3)を見ると1番目の“カン・ウォンレ”(姜元来)は韓国の音楽グループ・CLON(クローン)のメンバーの名前である。2番目と4番目は映画「スターウォーズ エピソード2:クローンの攻撃」と関連している。

⁷ 検索語には“クローン”(日)、“克隆”(中)、“클론”(韓)、“cloning”(英)を用いた。

Table 2 “クローン”の共起語（日本語）

No.	単語	記事数
1	病	222
2	人間	146
3	技術	149
4	携帯	56
5	体	78

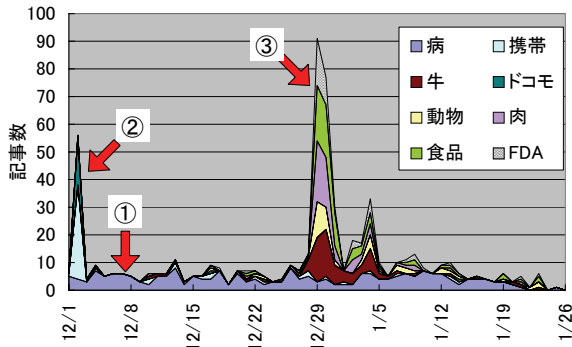


Fig. 4 “クローン”の共起語（日本語）
(積み上げグラフ, 主要語を抜粋)

Table 3 “클론”（クローンの音読み）の共起語（韓国語）

No.	単語	日本語訳	記事数
1	강원래	カン・ウオンレ	27
2	습격	襲撃	20
3	노래	歌	14
4	스타워즈	スターウォーズ	13
5	구절	句節	11

またFig. 5にこの時期の主な話題を表す共起語の推移を示す。図中、①はスターウォーズに対する関心、②はカン・ウオンレが2度目の交通事故にあったことに対する関心、③はカン・ウオンレが事故後、車椅子に乗り大学で講義を行ったことに対する関心であった。

第2に、意識である“복제”（複製）の共起語をFig. 6に示す。この時期、韓国ではES細胞事件の渦中にあったイ・ビョンチョン(李柄千)教授のソウル大学復職(図中①)や、狂牛病耐性牛クローンの誕生(図中②)、ファン・ウソク(黄禹錫)教授の研究再開に関する話題(図中③)に対する関心が見られた。日本語における関心と異なっていることが分かる。

“クローン”に関する日韓の関心比較

日本語と韓国語では“クローン”に対する関心が異なることが分かった。日本語圏ではクローン病やクローン携帯、クローン食品への関心が見られるのに対し、韓国語では映画や音楽グループ、またクローン研究に関連してイ・ビョンチョン教授やファン教授への関心が見られ、両者の関心は異なる方向を向いていることが分かる。

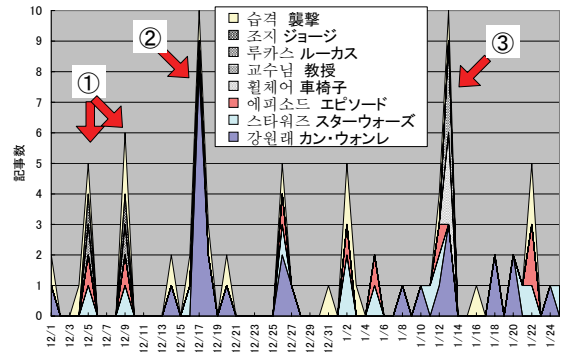


Fig. 5 “클론”（クローンの音読み）の共起語（韓国語）
(積み上げグラフ, 主要語を抜粋)

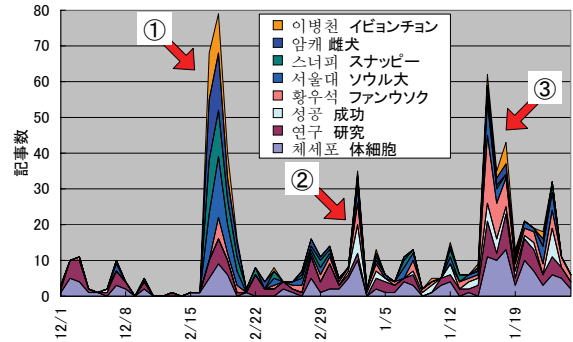


Fig. 6 “복제”（複製）の共起語（韓国語）
(積み上げグラフ, 主要語を抜粋)

左記の日韓の関心の違いは、ブログ執筆者層の違いにも起因すると考えられる。つまり日本語圏では幅広い年齢層がブログを執筆しているのに対し、韓国語圏ではブログを執筆する年齢層が若年層に集中する結果、映画や音楽に対する関心が目立ったのではないかと筆者らは考える。今後、各言語においてブログ執筆者層の把握や推定について別途調査を検討する。

3.2 韓国語記事についての定性分析

ここでは韓国語を母語とする調査者に日本語と韓国語のブログ記事を読んでもらい、日本と関心の異なる話題について定性的分析を行った。ここでは“整形手術”を題材とした。整形手術は韓国国内において手軽に行われており、例えば大学に合格した娘のために親が二重瞼の整形手術をプレゼントしたり、逆に子供が母親のしわを取るために整形手術をプレゼントすることが行われている。

ここでは(1)“男子整形”(남자 성형)と(2)“美顔手術”(얼굴 수술)というキーワードで日韓のブログ記事を検索し内容を分類した。検索期間は2006年9月1日から12月22日まで(16週間)、ヒット件数は韓国語176件(52件(1)+124件(2))、日本語150件(33件(1)+117件(2))であった。Fig. 7に各キーワードの記事数比較を示す。(2)については日韓とも同じ量の記事数であるが、(1)については韓国語記事の方が多い。

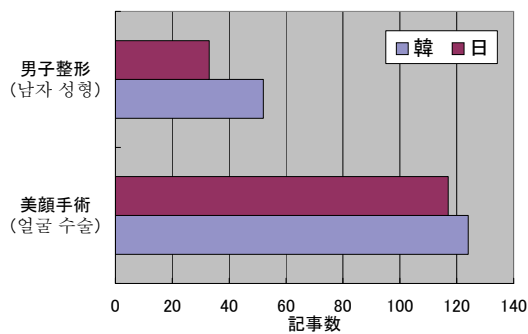


Fig. 7 整形手術に関する日韓比較

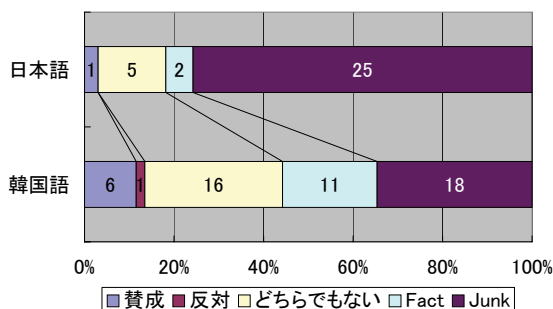


Fig. 8 “男子整形”に関する記事内訳 (図中数値は記事数)

次に検索結果の記事を手作業で5つのカテゴリに分類した。用意したカテゴリは1.賛成, 2.反対, 3.どちらでもない, 4.事実 (Fact), 5.左記以外 (Junk)である。分類結果を Fig. 8, Fig. 9 に示す。Fig. 8, Fig. 9 とも日本語では Junk の割合が高い。日本では整形手術に関する賛否の記事は殆ど無く、代わりに整形手術を薦める宣伝記事が多く見られた。これに対し韓国語では Junk の割合は日本語に比べ低く、賛否に関する記事も1割ほどあることが分かった。韓国語のブログの中には、韓国人の10代の女性の半数が整形を希望しているとの記述も見られた。また韓国語記事の場合、検索語(1)および(2)について「どちらでもない」が全体の3割を占めたが、その大半は整形を肯定するもので、最後に副作用の事例を紹介する内容であった。これらの結果、日本と韓国で整形手術に対する関心ならびに考え方の違いを垣間見ることができた。将来的にはこうした言語間における視点の相違を自動的に抽出したり可視化することを目標とする。

4. まとめと今後の展望

本論文では複数の言語で記述されたブログ記事から関心動向の比較と解析を行う言語横断型関心解析システムの構想と実装システムについて述べた。提案システムを用いて日韓のブログ記事を対象とした定量・定性分析を行い、日韓における関心の方向や考え方の違いについて報告した。本研究の関連研究には GALE プロジェクト⁵⁾がある。GALE プロジェクトは各国語で記述された文書に機械翻訳や情報抽出、自動要約等の自然言語処理を適用し、情報分析者を支援する構想である。多様な言語を扱う点で本研究と GALE は関連するが、

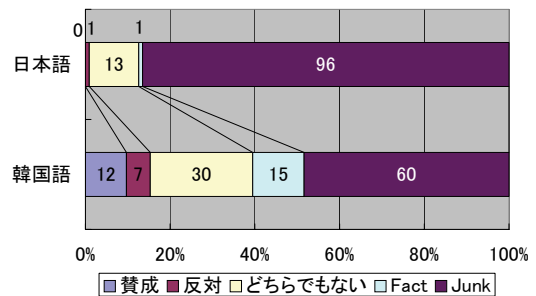


Fig. 9 “美顔手術”に関する記事内訳 (図中数値は記事数)

本研究では世界中の膨大なブログ記事からのボトムアップな関心動向の把握を目標としている。なお、即時性や解析速度については今後いっそうの改善を図る予定である。

また、今後、言語横断型関心解析のタスクにおいては、定量的な評価法を確立することが不可欠であるが、本研究においては、以下の方針に基づいて評価法を確立したい。まず、ベースラインとしては、アクセス可能な全ての外国語文書の集合全体を、その言語の母語調査者に和訳させ、その結果から価値の高い情報を収集する過程を設定し、その際に必要となる翻訳コスト、情報収集コスト、および、収集される情報の量を基準とする。ここで、言語横断型ウェブログ関心解析タスクにおいて解決すべき課題は、人間の翻訳者が翻訳するに足る価値ある文書を、膨大なブログ記事から絞り込むことである。そして、この過程において、ベースラインと比べて、(a)どの程度手動翻訳に要するコストを削減できるか、また、和訳された結果から、(b)どれだけ少ないコストで、(c)どれだけ多くの情報を収集できるか、といった項目を評価尺度と考えて、定量的な評価を行う。また、この定量的な評価においては、翻訳ソフトや対訳辞書、多言語間での統計的対照分析技術等を活用して、その有用性を検証する。

謝辞

本研究は科学技術振興機構・社会技術研究開発センターの支援によって行われた。また本研究の遂行にあたっては株式会社ナビックス・村上様、河田様との議論が有益でした。ここに謹んで感謝致します。

参考文献

- 1) 福原知宏, 中川裕志, 西田豊明: 時系列テキスト集合からの社会的関心の分析, 第16回インテリジェント・システム・シンポジウム予稿集, 1B1-3 (2006).
- 2) 武田英明, 大向一輝: Weblog の現在と展望—セマンティック Web およびソーシャルネットワークワーキングの基盤として, 情報処理, Vol. 45, No. 6, pp. 586-593 (2004).
- 3) Zhang, H.P., Yu, H.K., Xiong, D.Y., and Liu, Q.: HHMM-based Chinese lexical analyzer ICTCLAS, In Proceedings of Second SIGHAN Workshop on Chinese Language Processing, pages 184-187(2003).
- 4) Tsuruoka, Y. and Tsujii, J.: Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data, Proceedings of HLT/EMNLP 2005, pp. 467-474 (2005).
- 5) GALE Project: <http://www.darpa.mil/ipto/programs/gale/>